Drue, C. R. & Moser Deegan, M. (2026). Exploring Generative AI to Enhance Data Artifact Use in Accreditation. *Intersection: A Journal at the Intersection of Assessment and Learning, Early View*.

Exploring Generative AI to Enhance Data Artifact Use in Accreditation

Christopher Drue Ph.D. & Michele Moser Deegan Ph.D.

Author Note

Christopher R. Drue, https://orcid.org/0000-0002-0678-2336

"We have no conflicts of interest to disclose"

Intersection: A Journal at the Intersection of Assessment and Learning

Early View

Abstract: This study explores the viability of generative artificial intelligence (AI) tools to evaluate institutional documentation for alignment with accreditation standards. Using a portion of an institution's evidence inventory, researchers compared manual review with AI-supported methods, including a proprietary beta tool and a publicly available large language model (LLM). Employing a deductive framework informed by Constant Comparative Analysis (CCA), as well as a quantitative comparison, the study assessed AI's ability to place evidence in the correct standards and criteria required by the accreditor. Initial tests across 17 criteria, nested within three standards, revealed limited accuracy in category identification by LLMs. However, when the task was simplified to a broader set of standards, performance improved significantly, demonstrating LLM utility as a preliminary sorting aid. While easily accessible LLMs lack the precision and scalability required for full automation, findings point to promising directions for future development of tailored AI solutions to support accreditation.

Keywords: *higher education accreditation, generative AI, large language models*

Introduction

Higher education reaccreditation requires institutions to perform a self-study approximately once every 8-10 years as a key component of this process. To create the Self-Study report, institutions collect and analyze a myriad of documents and data on students, faculty, and staff, and other institutional evidence such as reports, communications, and policies, against accreditation standards. The accumulation of this information, often referred to as an evidence bank, into a comprehensive and useable repository, is time-consuming and challenging, particularly at larger institutions that span multiple locations and hundreds of academic and administrative offices. Until recently, there has been limited information about the preferred documentation that institutions should analyze and use for the Self-Study, making the process of evidence gathering more stressful for those responsible for the creation of the evidence inventory as the wrong evidence could trigger additional analyses and data collection or inadvertently flag a problem area because the evidence used in the report was incorrect or incomplete. Recognizing this challenge, in 2022 the Middle States Commission on Higher Education (MSCHE) approved its 14th edition of the *Standards of Accreditation and Requirements of Affiliation* to assist institutions with this information. In addition to refining the Standards for Accreditation (Standards), they created a document, *Evidence Expectations by Standard Guidelines* (Guidelines), to

assist institutions in collection and analysis of evidence that MSCHE believes is useful to successfully complete the Self-Study (Middle States Commission on Higher Education, 2023a; 2023b).

Over the past year, our institution's Self-Study Evidence Inventory Committee has undertaken a comprehensive effort to collect and organize documentation in preparation for the upcoming MSCHE Self-Study. Informed by the document, *Guidelines*, the committee curated approximately 800 discrete pieces of evidence, now housed within a cloud-based, vendor-provided data lake. These artifacts, used as evidence, span multiple formats including PDF, Excel, and Word documents, representing a broad cross-section of institutional operations, policies, and outcomes.

Prior to the inception of the Self-Study process, a critical question emerged: Do we possess the "right" evidence for each element outlined in the evidence document? Beyond quantity, we needed to consider the sufficiency and relevance of each document to ensure comprehensive coverage and alignment with MSCHE's seven standards and their associated criteria, which further refine the focus of inquiry. Given the scale and complexity of this task, our team explored the use of artificial intelligence (AI) to assist in identifying evidence gaps and validating the completeness of our inventory.

This study investigates the potential of AI to support institutional effectiveness and accreditation readiness by comparing traditional manual auditing methods with two AI-driven approaches: (1) the beta-version AI embedded within our assessment platform, and (2) a publicly available large language model (LLM) generative AI tool. Using a curated subset of our evidence inventory, we conducted a comparative analysis to evaluate how effectively these technologies can (a) identify missing or misaligned evidence, (b) support mapping to MSCHE Standards and Criteria, and (c) enhance the efficiency and accuracy of compliance documentation. We pose these research questions:

- To what extent can AI tools accurately identify and align institutional evidence with MSCHE Standards and Criteria compared to a manual review?
- What types of evidence are most frequently misclassified by AI, and what patterns emerge across content types?
- How might AI-assisted evidence mapping improve the efficiency, consistency, and completeness of institutional compliance documentation for accreditation purposes?

The study has wide applicability, not just for institutions engaged in Self-Study, which is a required element of nearly all institutional and professional accreditors. It is also applicable for other types of continuous improvement reviews including academic assessment and program review, as well as performance and financial audits or strategic planning. Organizations beyond higher education may also benefit from this type of AI use case to match documents to processes and outcomes.

Literature Review

Higher Education Accreditation

Broadly speaking, the accreditation process has evolved from a mechanism primarily for internal improvement and public health protection in the 1890s towards a tool used for external compliance and accountability (Brittingham, 2009; Ewell et al., 2009, Eaton 2010). While modifications to the higher education accreditation landscape continue, in the United States, the core goal of these

processes, to assure educational quality and confidence in educational outcomes, remain constant (Brittingham, 2009; Ewell et al., 2009, Nguyen et al., 2021). The current system of accreditation of higher education institutions in the United States is unique among nations. Institutional accreditation is voluntary. Institutions may choose not to be recognized by an institutional or professional accreditor, although without institutional accreditation, they are ineligible for financial aid funds for students and peer review of accountability metrics. The process for management of accreditation is also unique as quality assurance is managed through a Triad governing structure that includes the federal government, federally recognized accreditors, and state governments. The federal government's primary responsibilities are to oversee financial aid funds and federally recognized accreditors. The federal government does not directly accredit institutions, rather, it relies on federally recognized institutional accrediting organizations, such as the Higher Education Learning Council and Middle States Commission on Higher Education, which accredit an entire institution, and programmatic accreditors, such as the American Bar Association or American Medical Association, which accredit programs, typically within larger institutions (Brittingham, 2009). Higher education institutions and programs voluntarily agree to standards set by these federally recognized self-regulating accrediting agencies that monitor institutional activities to ensure compliance with these standards (Eaton, 2010). Accrediting organizations rely on voluntary support from members of accredited institutions to carry out the tasks of accreditation such as site visits and reviews of substantive changes to institutional academic program offerings.

While higher education institutions in the U.S. have existed since the 17th Century, federal government involvement to regulate the system was slow to emerge, leaving space for the higher education community and state governments to develop accreditation practices without federal governmental involvement (Brittingham, 2009). Federal involvement in accreditation can be traced back to the 1940s, with the passage of the Servicemen's Readjustment Act (1944) and National Defense Education Act (1958), as the law required that beneficiaries of the program's tuition benefits could only apply federal funds to cover the cost of education at accredited institutions. Between 1950 and 1965, accrediting agencies adopted what are considered today's fundamentals in the accreditation process: a standards-based self-study and a review and site visit by voluntary peers in higher education (Eaton, 2010).

The Higher Education Act (HEA) of 1965, and subsequent reauthorizations have further elevated the importance of accreditation, ruling that federal student funding opportunities are only available to institutions accredited by a US Department of Education-approved entity (Eaton 2010; Romanowski & Karkouti, 2022). By the mid-1980s, the call by federal policymakers for increased accountability across all education levels provided more explicit directives to accrediting agencies. Accrediting agencies responded to this call for action with new guidelines for accreditation and greater professional support for institutional staff responsible for managing assessment and accreditation (Brittingham, 2009). Higher education leaders responded to these events by creating new professional organizations and conferences, such as the Assessment Forum of the American Association for Higher Education and expanded the scope of existing organizations such as the Association for Institutional Research, evolving the field of institutional research. The emerging culture of continuous improvement was bolstered by computing capabilities and the expansion of personal computers in the workplace,

increasing the capacity to apply more sophisticated statistical methods to measure student learning and other outcomes (Brittingham, 2009; Ewell et al., 2009).

As mentioned previously, while the federal government does not directly accredit higher education institutions, since the 1992 HEA reauthorization, the National Advisory Committee on Institutional Quality and Integrity (NACIQI), has added an additional level of federal oversight over accreditation. This council advises the U.S. Secretary of Education on matters of accreditation including the recognition process for accrediting agencies and recommendations for degree approvals. It also monitors and reports on educational quality (NACIQI, 2025). These recommendations have resulted in federal regulatory change, which impacts policies and practices of the accreditors and accredited institutions or can result in termination of an accrediting agency (Kelchen, 2017).

The Spelling Report and subsequent passage of the Higher Education Opportunity Act in 2008 refocused attention on accountability and quality assurance. The Act reconstituted NACIQI, provided additional direction for accountability, including more transparency regarding cost, price, and student success outcomes, and an institution's "value-added" (Spelling Report, 2006). The report was issued at a time when state investment in higher education was decelerating, limiting robust state responses; however accrediting agencies responded to the report by renewing standards and creating new monitoring policies and procedures for accredited institutions (Ewell et al., 2009).

The evolution of tools for institutional assessment, such as the AAC&U VALUE rubrics (Association of American Colleges and Universities, n.d.), helped institutions adapt to accountability changes. Technological advances in systems designed to manage large and complex data sets, "big data" also emerged during this period alongside the creation of cloud-based storage systems by Google and Microsoft. As accreditation standards changed and the accountability movement grew, institutional tension increased between those arguing that these efforts limit academic freedom (Holt, 2020; Romanowski & Karkouti, 2022) and those concerned about assessment's growing use for external accountability (Ewell et al., 2009).

Despite efforts by the higher education community to bolster accountability through accreditation, in 2020, the federal government altered the accreditation landscape. The administration's goals for these changes were improved student outcomes, return on investment, and speeding up the time it takes for an institution to receive accreditation (Fain, 2019). The primary result of these efforts was to remove geographic boundaries from regional accreditors, which are now known as institutional accrediting bodies, increasing competition across accreditors. While some higher education institutions took advantage of the opportunity to become accredited by an organization outside of their geographic region, this regulatory change did not sufficiently meet the higher education goals of President Trump (Lederman, 2020). Returning to office in 2025, the administration has engaged in additional administrative reform, including an Executive Order designed to accelerate attention toward, "high-quality valuable education for students," (United States Government, Executive Office of the President, 2025). In June 2025, Governor Ron DeSantis (FL) announced a new institutional accrediting body, The Commission for Public Higher Education, which is seeking recognition by the USDOE, following several states' dissatisfaction with their previous accrediting body. Additionally, the Administration directed the Attorney General and Secretary of Education to investigate accrediting agencies to ensure lawful

practices of the entities and institutions that they accredit. As the accreditation environment continues to evolve, institutions face increasing pressure to produce measurable outcomes and maintain efficient, well-structured assessment systems that can withstand heightened scrutiny.

Amid the current political challenges to accreditation systems and processes, AI is creating changes to the ways in which those involved in accreditation manage data and documents, with the potential to advance accreditation practices to meet the increasingly complex accountability environment. As AI matures for this purpose, it can be used as an accelerant to quickly analyze qualitative and quantitative sources to meet internal and external mandates. While not calling out AI directly in the April 23, 2025 Executive Order, the document provides a directive to accreditors to, “increase the consistency, efficiency, and effectiveness of the accreditor recognition review process, including through the use of technology,” opening the possibility of the robust use of AI for accreditation (United States Government, Executive Office of the President, 2025).

AI in Higher Education

Discussions surrounding AI permeate higher education, and it has become an emerging tool for experimentation in assessment and accreditation. Perhaps most pressingly, students' use of generative AI on formative and summative assessments has upended the nature of assessment in higher education. In response to the surge in use of AI, scholars have called for curriculum reform to ensure that students are learning the outcomes that are necessary for their education (Bennett & Abusalem, 2024). Concerns about the negative consequences of using AI for educational purposes have been balanced by the recognition that students also need to be prepared to enter a workforce where AI may be part of everyday activities (Joshi, 2025).

Apart from its impact on student learning, AI also presents opportunities for faculty and administrators to boost productivity, particularly by automating repetitive and time-consuming tasks used in assessment. For example, Slotnick and Boeing (2024) conducted a study using generative AI to interpret program-level assessment data. They did initial familiarization and development of the codebook using a human-only approach and then incorporated common LLMs to assist in categorizing the full dataset. While the tools and prompts they used made errors along the way, they concluded that judicious implementation could make AI a "valuable tool" for this type of work (Slotnick & Boeing, 2024, p. 14).

Others have argued that AI can play an even broader role in assessment practices, helping instructors design assessments and innovative test questions, and providing formative and summative feedback to learners (Trajkovski & Hayes, 2025). Such capabilities suggest a transformative shift in how assessments are conceptualized and implemented, potentially leading to more equitable and effective educational outcomes. As an example of how this might work, DeSabito et. al. (2024) developed a custom (and proprietary) AI model using ChatGPT to assess learning outcomes in samples of student papers. The AI model, which they named "Walter," was trained using rubrics to evaluate student work, and the scores were compared to human evaluators, with mixed results. While the AI model and human evaluators rated students similarly on some types of learning outcomes, they found large differences in others. They concluded that the AI-human partnership for assessment worked best when

assignments were carefully crafted to be used across courses and disciplines and when assessment rubrics concretely identified desired student output (DeSabito et. al., 2024, p. 12).

Beyond assessment, the emergence of readily available and low cost LLM AI platforms has prompted scholars to explore the effectiveness of using such tools for qualitative analysis. Qualitative analysis is the process of coding textual evidence into themes and comparing the context and frequency of these themes. By using appropriate sequences of prompts, even free versions of tools such as ChatGPT can productively sort short qualitative datasets into themes (Zhang et al., 2025). However, these tasks require an iterative approach to prompting and constant checking by informed researchers, and the AI program cannot replace human interpretations of the results. As Zhang and colleagues wrote, "treating ChatGPT solely as a tool may lead to over-reliance, where researchers undervalue critical engagement with the AI's outputs" (Zhang et al., 2025, p. 12). But others warn that the limitations of using generative AI for qualitative research, from AI fabrication to biases, make it less clear that AI will ever do an adequate job of conducting qualitative research (Roberts et al., 2024).

While AI has transformative potential in assessment, integrating AI into course and program level assessment and related research tasks raises important questions about data privacy, algorithmic bias, and the essential role of human judgment. It is also a complex process to develop custom AI applications that are focused specifically on educational assessment tasks.

Case Study Description

The case study for this quasi-experiment is a comprehensive public research university accredited by the Middle States Commission of Higher Education (MSCHE), with an enrollment of over 70,000 and multiple educational sites throughout the state. With a complex administrative structure and over 150 undergraduate majors and 400 graduate programs, the committee charged with gathering evidence for our reaccreditation needed to ensure representation of the varied experiences and administration of policies and processes that reflect the diversity of our programs and locations, while meeting the data and evidence requirements of our accreditor.

New to the *Standards for Accreditation and Requirements of Affiliation*, 14th Ed., MSCHE makes available a document, Evidence Expectations by Standard Guidelines, which provides examples of evidence that they recommend or require institutions make available to support the written Self-Study. This document is formatted by standards, with criteria to further clarify the intentions of each standard. Examples of evidence are provided for each criterion. For example, Standard I, Criteria 1 focuses on mission and goals. The suggested or required evidence to correspond to this criterion include mission statements and related evidence, sample communications related to mission statements, evidence of alignment between the mission statement and planning, resource allocation, etc., and sample budget requests or other documentation demonstrating alignment. As part of the Self-Study, MSCHE recommends that institutions undergoing accreditation or reaccreditation populate an evidence bank with examples and cite these examples in the written Self-Study report.

In preparation for the creation of the Self-Study, the evidence inventory committee used the abovementioned document to gather several hundred documents into a cloud-based data solution,

which is used to prepare the Self-Study report draft and designed to more easily link evidence to the appropriate sections of the draft. The vendor hosting the cloud-based system recently added an AI pilot to their platform, and we were asked to pilot it for use with our Self-Study. This pilot created the opportunity for a natural experiment, comparing the cloud-based in-house AI system to other AI tools.

Method

This study employs a deductive and artificial-intelligence-aided application of Constant Comparative Methodology (referred to as CCM or CCA) (Boeije, 2002). While CCM is generally used as an inductive approach, a deductive application of its principles can be used to systematically analyze qualitative data when a pre-existing theory or conceptual framework guides the research. In this approach, instead of allowing categories to emerge purely from the data, we begin with a set of *a priori* codes derived from the accreditation guidelines. The process then involves constantly comparing new data segments (or files containing evidence) to these pre-defined categories, assessing how well the evidence fits, supports, or challenges existing constructs. This iterative comparison helps to confirm, refine, and even identify limitations of the initial categorization, making it a powerful method for exploring our accreditation artifacts, albeit distinct from the theory-generating goal of traditional grounded theory.

While CCM ensured a rigorous approach to our initial categorization, we also incorporated a quantitative evaluation process from the machine learning domain to assess the effectiveness of AI in replicating human-coded insights. To compare results, we present the weighted precision (ratio of true positives to the sum of true positives and false positives), weighted recall (ratio of true positives to the sum of true positives and false negatives), and the weighted F1 score for each of the tests. The F1 score is a type of F Measure that is often used to compare the results of categorization tasks using machine learning, and it is also a useful way to assess LLM categorization since it offers a balanced metric showing the accuracy of a categorization task (Than et al., 2025; Thelwall, 2022). This mixed-methods approach not only preserves the richness of the qualitative coding process but also demonstrates the scalability and reliability of AI assisted coding.

The F1 score is defined as the harmonic means of precision and recall. It is especially valuable in contexts with uneven class distribution, such as accreditation evidence mapping, where individual artifacts typically align with only one or two criteria. In such cases, a balanced consideration of precision and recall becomes critical for fair and effective categorization. The weighted F1 score is the weighted average of the F1 scores for each class, where the weight is determined by the number of instances in each class (Christen et al., 2024). The F1 score is an imperfect measure, designed to ignore true negatives, or cases in which irrelevant sample cases are correctly identified as not belonging to any of the categories (Christen et al., 2024). Nonetheless, given that it is an accepted measurement strategy, scholars have generated standards for acceptable scores that are comparable to inter-rater differences between two human coders.

For this study, we categorize evidence documents based on MSCHE Standards and Criteria and compare the effectiveness of various commonly available AI platforms to human experts. To ensure the rigor and consistency of this process, all our comparison cases use the same detailed codebook

that clearly defines each standard and criteria with illustrative examples. The MSCHE document, *Evidence Expectations by Standard Guidelines* (Middle States Commission on Higher Education, 2023a) serves as this "codebook."

For the human-controlled portion of this research process, we established inter-rater reliability by having each researcher independently review and categorize every document in the sample according to MSCHE Standards and Criteria. Following individual coding, the researchers met to reconcile discrepancies and reach consensus on differing interpretations of the standards. While many artifacts may tangentially relate to MSCHE categories, superficial keyword matching is insufficient for meaningful classification. Such an approach risks producing a fragmented and incoherent body of evidence for self-study. Instead, the research team applied a more rigorous criterion: an artifact was coded to a standard only if a professional in institutional assessment would reasonably identify and use it as evidence in support of that standard.

To validate our initial coding decisions, we conducted a secondary review of all AI-generated categorizations. Each AI output was carefully examined to ensure that any omissions in our original coding did not inadvertently influence the model's results. This process was both labor-intensive and time-consuming, particularly within the context of assessment workflows, which underscores our motivation for exploring AI-assisted solutions.

The treatment is the comparison of the control to the AI platforms. We ran our test of AI systems using several models to explore the capabilities of easy-to-access systems for cataloging assessment evidence. We chose our LLM AI systems based on our university's data classification and AI policy, which identified platforms that were secure for use with different types of internal data. In our case, two tools were approved for the type of evidence most common in accreditation work.

The first AI platform we selected was Microsoft Copilot. This platform is available to faculty and staff on our campus and is approved for use by our data security office for the type of internal planning documents and public reporting documents that we used for this test. Unlike other AI systems with publicly versioned releases, Copilot operates without public version identifiers. For this study, we used the version available through our university subscription during August and September of 2025. As Copilot updates continuously, this timeframe serves as the basis for describing its capabilities and outputs. We chose this system because it is widely available (comparable to ChatGPT), though there are limitations such as restrictions on the number of files that can be uploaded at a time. Using this platform is most accessible for other assessment offices, and so we felt it was important to test this solution. The second AI platform we selected was a custom vendor-provided AI tool, powered with ChatGPT, that is integrated into our assessment and accreditation platform. This custom AI is restricted to reviewing the evidence that is uploaded into the system, but it was not otherwise trained on assessment data.

We chose a zero-shot prompting strategy because we were working with chatbots that had limited context windows, and we did not have the resources to develop a custom AI tool for this task. Zero-shot prompting is when an AI model is given a task with no prior examples or training on that specific task, relying solely on the prompt wording to generate a relevant response. A zero-shot strategy is

most realistic in mimicking how assessment professionals would likely utilize an LLM AI platform to parse accreditation evidence. We also varied our approach by starting "new conversations" between each subsequent prompt for some tests, while retaining the system "memory" throughout the test, to the extent possible, in other cases.

We tested a variety of prompts, considering the length of the prompts, the information requested, and temperature threshold (certainty) required for the categorization. Focusing on MSCHE's Standard I: Mission and Goals, Standard III: Design and Delivery of the Student Learning Experience, and Standard V: Educational Effectiveness Assessment, the MSCHE codebook was used as part of a prompt for the AI platforms by selecting each standard and criteria and creating a separate line that concatenated the standard with the description and examples of evidence. For each platform, we began with a testing phase where we tried several iterations of prompts and methods of providing the codebook and files to the platform. We settled on the prompts that performed best at the task and provided relevant portions of the final prompts in Appendix A. The full prompts are quite long, ranging from 700-4,000 words, because the MSCHE codebook is detailed and extensive. As part of our prompting strategy, we developed shorter prompts (Appendix A: CoPilot Prompt E, Vendor AI Prompt B) using the assistance of the LLM to summarize the MSCHE Standards.

Sample

A sample of artifacts was collected from the evidence bank and used for testing. The sampling method was informed by our institution's data classification and AI policy, which describes the types of information that may be used in university-approved AI tools. Following this guidance, we selected items of the following types, all which were approved for use in our institutionally licensed tools (Co-Pilot and our vendor-provided assessment tool):

- Publicly available reports and data and communications, including information that is collected and published by the federal government, and
- Materials intended for publication and shared on non-password protected university webpages such as catalogs, student and staff handbooks, and internal policy and procedure documents.

We collected the sample by reviewing our evidence bank and choosing examples of each type of artifact for each standard, while excluding evidence that contained sensitive information. We then added four additional artifacts that could not be plausibly tagged to any of the three standards. This resulted in a sample of 53 artifacts. Our instance of CoPilot had a 100-page limit and 50mb size limit for files, so we edited the few files that were too large by removing pages or images until the files were acceptable. In all cases, resizing the artifacts did not remove critical information for the classification process. The artifacts were given descriptive titles that did not refer to the MSCHE standards so as not to influence the coding results.

As noted above, the selection of evidence included internal strategic plans, institutional websites, reports that are available on public websites, internal assessment summaries, and IPEDS reports. While a complete set of evidence would be valuable, limiting our sample was necessary to compare the AI models. The resultant sample provided ample opportunity to assess the capabilities of the tools and

were not substantially different in nature from the artifacts that were excluded through the selection criteria.

Analysis

Table 1, below, presents the weighted precisions, weighted recall, and weighted F1 scores for four initial tests. Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive. It answers the question: "Of all the items the model predicted were relevant, how many actually were?" A model with high precision makes very few false positive errors. Recall measures the proportion of actual positive instances that were correctly identified by the model. It answers the question: "Of all the items that were actually relevant, how many did the model find?" (Than et al., 2025; Thelwall, 2022). A model with high recall makes very few false negative errors. Precision and recall are both calculated based on formulas that yield values between 0 and 1, with 1 denoting perfect prediction. The F1 score is the harmonic mean of precision and recall and also yields a value between 0 and 1. While there is considerable flexibility in interpreting the value of the F1 score, generally values above 0.6 are interpreted as moderate agreement, while values of 0.8 are interpreted as showing high agreement between a model's output and human reviewers (O'Connor & Joffe, 2020; Than et al., 2025).

Table 1

Weighted Performance Metrics Across AI Models Evaluated on 17 Accreditation Criteria

	CoPilot Test 1	CoPilot Test 2	CoPilot Test 3 - Short Prompt	Vendor Platform Test 1	Vendor Platform Test 2 - Short Prompt
Weighted Precision	0.561	0.762	0.778	0.443	0.296
Weighted Recall	0.825	0.439	0.658	0.465	0.462
Weighted F1 score	0.653*	0.526	0.682*	0.443	0.353

Note: Cells shaded in blue, also denoted with an asterisk, have a weighted F1 score of 0.6 or higher.

Table 1 shows the results of our first series of tests, in which our AI platforms were prompted to determine whether each artifact was a "clear match" or "no match" for each of the 17 criteria across the three standards. The variation in the table between each test reflects the reality that LLMs do not produce the same output for different users or even for the same users at different times. This variability stems from the way large language models (LLMs) operate. LLMs are built on deep learning architecture and trained on massive datasets to predict the next word in a sequence based on context. Their outputs are probabilistic rather than deterministic (Minaee et al., 2024).

Our first research question asked to what extent AI tools can accurately identify and align institutional evidence with the three MSCHE standards and 17 criteria compared to manual review. The weighted

F1 scores between 0.3 and 0.6 indicate substantial errors in evidence classification. Across 17 criteria, the AI LLM platforms we tested struggled to accurately identify the correct matrix of applicable categories. This task requires nuanced interpretation, as the overall standard is written broadly, while its criteria are specific and pointed. For example, Standard I: Mission and Goals Criteria 1 is as follows: "A candidate or accredited institution possesses and demonstrates the following attributes or activities:

1. clearly defined mission and goals that:
    a. are developed through appropriate collaborative and inclusive participation by all who facilitate or are otherwise responsible for institutional development and improvement;
    b. address external as well as internal contexts and constituencies;
    c. are approved and supported by the governing body;
    d. guide faculty, administration, staff, and governing structures in making decisions related to planning, resource allocation, program and curricular development, and the definition of institutional and educational outcomes;
    e. include support of scholarly inquiry and creative activity, at levels and of the type appropriate to the institution;
    f. are publicized and widely known by the institution's internal stakeholders;
    g. are periodically evaluated." (Middle States Commission on Higher Education, 2023b)

Unlike AI, human assessment professionals rely on their expertise to select the most compelling evidence for accreditation reports. They understand what a review committee finds convincing and can distinguish between mere mentions of a statistic (e.g., in a promotional pamphlet) and systematic evidence of measuring student achievement.

Since the LLM platforms have limited context windows, we explored the difference between longer prompts that contained the complete MSCHE Criteria, which were approximately 4000 words, and shorter prompts that provided an AI-abbreviated criteria and came in at fewer than 1000 words per prompt. We created these shorter prompts by asking each LLM to shorten the full guide into key essential components. The results show that in one test (CoPilot Test 3), the shorter prompt performed marginally better than our best long prompt model, while the vendor platform performed worse. While the shorter prompt was easier to manage, it did not make substantial improvements to the AI classification of artifacts.

Given the challenge of the task, we can compare the results to an easier task: In the next test, we used the same underlying data, but merged the criteria so that we were only categorizing to MSCHE Standards I, III, and V. If the model correctly placed the item in a standard, regardless of whether it chose the correct criteria, we identified it as a true positive. If the model did not identify any criteria when the artifact matched at least one criterion, we identified it as a false negative. Similarly, if the model identified a match in any criteria when none fit, it was a false positive. Thus, we go from 17 total possible categories to 3 in this simplified test.

Table 2

Weighted Performance Metrics Across AI Models Evaluated on 3 Accreditation Standards

	CoPilot Test 1	CoPilot Test 2	CoPilot Test 3 - Short Prompt	Vendor Platform Test 1	Vendor Platform Test 2 - Short Prompt
Weighted Precision	0.870	0.953	0.914	0.889	0.773
Weighted Recall	0.943	0.686	0.701	0.575	0.651
Weighted F1 score	0.899*	0.793*	0.829*	0.692*	0.695*

Note: Cells shaded in blue, also denoted with an asterisk, have a weighted F1 score of 0.6 or higher.

As shown in Table 2, with this simpler configuration, performance of the AI was substantially improved, with CoPilot producing strong results. Used in this way, AI would allow for an accurate enough starting point for a busy professional to sort through a large quantity of digital evidence, using their professional judgement to make final decisions. The findings suggest a qualified affirmation to our first research question. While the tested LLMs demonstrated reasonable accuracy in identifying evidence aligned with overarching MSCHE Standards, they struggled to consistently and correctly associate evidence with the more granular criteria.

We were also interested in what types of evidence were most frequently misclassified by AI, and what patterns emerged across content types. To facilitate this analysis, the evidence inventory was classified into four functional categories based on content type, institutional purpose, and relevance. The first category, Assessment & Learning Outcomes Reports, includes artifacts such as annual assessment reports from schools and colleges, academic program review documentation, and templates for assessment. The second category, IPEDS & Institutional Data Submissions, encompasses federally reported datasets related to enrollment and library services. The third category, Strategic Plans, Mission Statements, and Vision Documents, includes strategic plans, websites with the university's vision described, and other documents that offer insight into institutional priorities and planning processes. Finally, Academic Policies, Faculty Affairs & Curriculum Documentation contained documents such as faculty personnel policies, student policies, catalogs and curriculums, and faculty resources.

Table 3 shows the average weighted precision, recall, and F1 score, across all five models, for each of the four categories of evidence. Because our sample sizes were substantially reduced when subdivided into these categories, several classes had zero support; meaning no actual instances were present. In such cases, we assigned precision, recall, and F1 scores a value of zero to ensure that performance metrics reflect only those categories where the model had a meaningful opportunity to make correct predictions. This is standard practice in classification reporting.

Table 3

Weighted Performance Metrics Across AI Models Evaluated on 17 Accreditation Criteria, by Category of Evidence

	CoPilot Test 1	CoPilot Test 2	CoPilot Test 3 - Short Prompt	Vendor Platform Test 1	Vendor Platform Test 2 - Short Prompt
Assessment & Learning Outcomes Reports	WP: 0.743 WR: 0.902 WF1: 0.789*	WP: 0.893 WR: 0.457 WF1: 0.585	WP: 0.921 WR: 0.714 WF1: 0.781*	WP: 0.838 WR: 0.443 WF1: 0.555	WP: 0.713 WR: 0.557 WF1: 0.593
IPEDS & Institutional Data Submissions	WP: 0.735 WR: 0.500 WF1: 0.624*	WP: 0.550 WR: 0.500 WF1: 0.511	WP: 0.660 WR: 0.500 WF1: 0.544	WP: 0.600 WR: 0.400 WF1: 0.471	WP: 0.317 WR: 0.450 WF1: 0.357
Strategic Plans, Mission Statements, and Vision Documents	WP: 0.768 WR: 0.789 WF1: 0.759*	WP: 0.627 WR: 0.357 WF1: 0.426	WP: 0.783 WR: 0.429 WF1: 0.539	WP: 0.648 WR: 0.610 WF1: 0.598	WP: 0.869 WR: 0.262 WF1: 0.376
Academic Policies, Faculty Affairs & Curriculum Documentation	WP: 0.832 WR: 0.769 WF1: 0.748*	WP: 0.795 WR: 0.462 WF1: 0.545	WP: 0.767 WR: 0.795 WF1: 0.770*	WP: 0.448 WR: 0.385 WF1: 0.403	WP: 0.354 WR: 0.513 WF1: 0.413

Note: WP = Weighted precision, WR = Weighted recall, WF1 = Weighted F1 score; weighted by support for each category. Cells shaded in blue, also denoted with an asterisk, have a weighted F1 score of 0.6 or higher.

We found that, on average, the AI platforms performed best at categorizing the Assessment and Learning Outcomes reports and the Academic Policies, Faculty Affairs and Curriculum Documentation. It was less successful at categorizing IPEDS & Institutional Data as well as Strategic Plans, Mission Statements, and Vision Documents. This is likely because IPEDS data often lacks written context explaining its use, requiring a comprehensive understanding of the purpose of the data to properly categorize. The strategic plans and mission statements had the opposite challenge: they made sweeping claims addressing the entire scope of work at the university while even presenting limited data to support positions. Assessment professionals using AI to categorize work should be cautious when engaging with these types of artifacts, which may require human oversight to properly parse into appropriate categories.

Discussion & Recommendations for Practice

We agree with Slotnick and Boeing (2024) that AI could become a useful tool for assessment research but concur with DeSabito et al. (2024) that an AI-human partnership works best. Even if we look only at the most accurate prompt and model, the explorations presented in this study were not practical and efficient for large scale categorization of assessment data. Nor would these techniques replace the expertise of trained assessment professionals. But they do suggest an avenue for future custom AI tools that could be either built in-house or produced by vendors that would aid in the process in a meaningful way.

Among the challenges of conducting the tests in the way we did was the limited context window for the LLMs we were using and the ability to upload only one file at a time in Copilot. In our initial tests, we split the prompt into 4 separate queries due to the limited context window, which required several inputs for each artifact. Were we to widen the lens to look at all the MSCHE Standards and Criteria and all collected evidence at an institution, such a test would be immediately impractical. Another challenge was that it took a considerable amount of time for the AI models to process our requests, especially in our vendor provided instance of ChatGPT. Some queries took 20 minutes or longer to process.

To enhance efficiency in accreditation workflows, a custom large language model (LLM) could enable simultaneous analysis of multiple artifacts while incorporating deeper training beyond the initial prompt, embedding domain-specific assessment knowledge and contextual background. With appropriate access to archived accreditation self-studies, such a model could generate predictions informed not only by the standards themselves, but by the actual documentation submitted in prior reviews. Our second analytical comparison (Table 2) showed that even an untrained LLM is relatively accurate when associating artifacts with standards.

Institutions routinely collect documentation for a variety of operational and strategic purposes, storing these materials in centralized data repositories. An AI-enabled system could automatically categorize incoming documents as potential evidence for future accreditation self-studies, offering real-time suggestions aligned with relevant standards. Assessment professionals would ideally retain oversight, reviewing and confirming these classifications when the self-study process begins, thereby streamlining evidence collection and reducing the need for retrospective searches. A vendor-developed tool with these capabilities would represent a significant advancement for accreditation and assessment offices, offering scalable support to assist in initial evidence mapping.

Despite the challenges we faced in applying AI to these tests, including its limited accuracy in correctly identifying criteria, the exercise proved valuable. While the LLMs frequently misclassified content, their outputs prompted us to reconsider our initial coding decisions. Both researchers acknowledged that, during our first pass through the evidence, we had overlooked categories that were, in retrospect, clearly applicable. Although the process was time-consuming, it ultimately strengthened the quality of our analysis. That improvement, however, was contingent on our careful, manual review of each AI-generated categorization.

While assessment and accreditation remain time-intensive and laborious, these processes are ultimately meant to drive continuous improvement and better outcomes for our students. The central question is whether we can harness AI responsibly to reduce time burdens, enhance operational efficiency, and refocus our efforts on what matters most. Although the tools tested did not yield immediate gains in efficiency, the exercise itself improved classification accuracy—an outcome that may translate into stronger results over time. Importantly, these findings suggest that even current technologies, with thoughtful investment and iterative development, hold promise for advancing evidence classification and supporting more effective accreditation practices.

References

Association of American Colleges and Universities. (n.d.). VALUE rubrics. https://www.aacu.org/value-rubrics

Bennett, L., & Abusalem, A. (2024). Artificial Intelligence (AI) and its Potential Impact on the Future of Higher Education. *Athens Journal of Education*, 11(3), 195–212. https://doi.org/10.30958/aje.11-3-2

Boeije, H. (2002). A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality & Quantity*, 36(4), 391–409. https://doi.org/10.1023/A:1020909529486

Brittingham, B. (2009). Accreditation in the United States: How did we get to where we are? *New Directions for Higher Education*, 145(7). https://doi.org/10.1002/he.331

Christen, P., Hand, D. J., & Kirielle, N. (2024). A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. *ACM Computing Surveys*, 56(3), 1–24. https://doi.org/10.1145/3606367

DeSabito, D., Hansen, L., Mennella, T., & Rodriguez, J. (2024). Exploring the frontiers of generative AI in assessment: Is there potential for a human-AI partnership? *New Directions for Teaching and Learning*, 182, 81-96. https://doi.org/10.1002/tl.20630

Eaton, J.S. (2010). Accreditation and the Federal Future of Higher Education, Academe Magazine, American Association of University Professors, September-October.

Ewell, P. T., Kuh, G. D., & Ewell. (2009). Assessment, Accountability, and Improvement: Revisiting the Tension. National Institute for Learning Outcomes Assessment.

Fain, P. (2019). New Rules on Accreditation and State Authorization. *Insider Higher Education*, October 31.

Holt, B.J. (2020) Identifying and Preventing Threats to Academic Freedom, *International Journal for Infonomics*. 13(2), 2205-2012.

Joshi, S. (2025). Training US Workforce for Generative AI Models and Prompt Engineering: ChatGPT, Copilot, and Gemini. *International Journal of Science, Engineering and Technology*, 13(1), 1–11. https://doi.org/10.61463/ijset.vol.13.issue1.155

Kelchen, R. (2017). Higher Education Accreditation and the Federal Government, Urban Institute, September.

Lederman, D. (2020). Go East (or North), Regional Accreditor, *Insider Higher Education*, February 29.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large Language Models: A Survey. Cornell University Library, arXiv.org. https://doi.org/10.48550/arXiv.2402.06196

Middle States Commission on Higher Education. (2023a). Evidence expectations by standard guidelines: Aligned with the Standards for Accreditation and Requirements of Affiliation (14th ed.). https://www.msche.org

Middle States Commission on Higher Education. (2023b). Standards for Accreditation and Requirements of Affiliation (14th ed.). https://www.msche.org

NACIQI (2025). U.S. Department of Education National Advisory Committee on Institutional Quality and Integrity, accessed September 5, 2025. https://sites.ed.gov/naciqi/

Nguyen, C.H., Marshall, S.J. & Evers, C.W. (2021). Higher Education Quality Assurance and Accreditation Implementation in Several Countries across the World and Learned Lessons for Vietnam, *Vietnam Journal of Education*, 5(1), 11-17.

O'Connor, C., & Joffe, H. (2020). Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods*, 19, 1-13. https://doi.org/10.1177/1609406919899220

Roberts, J., Baker, M., & Andrew, J. (2024). Artificial intelligence and qualitative research: The promise and perils of large language model (LLM) 'assistance'. C*ritical Perspectives on Accounting*, 99, 102722. https://doi.org/10.1016/j.cpa.2024.102722

Romanowski, M. H., & Karkouti, I. M. (2022). United States accreditation in higher education: does it dilute academic freedom?. Informa UK Limited. https://doi.org/10.1080/13538322.2022.2127165

Slotnick, R. C., & Boeing, J. Z. (2024). Enhancing qualitative research in higher education assessment through generative AI integration: A path toward meaningful insights and a cautionary tale. *New Directions for Teaching and Learning*, 1-16. https://doi.org/10.1002/tl.20631

Spellings, M. (2006). A Test of Leadership: Charting the Future of U.S. Higher Education. Washington, DC: U.S. Department of Education.

Than, N., Fan, L., Law, T., Nelson, L. K., & McCall, L. (2025). Updating "The Future of Coding": Qualitative Coding with Generative Large Language Models. *Sociological Methods & Research*, 54(3), 849–888. https://doi.org/10.1177/00491241251339188

Thelwall, M. (2022). Can the quality of published academic journal articles be assessed with machine learning? *Quantitative Science Studies*, 3(1), 208–226. https://doi.org/10.1162/qss_a_00185

Trajkovski, G., & Hayes, H. (2025). AI-assisted assessment in education: Transforming assessment and measuring learning (1st ed.). *Springer Nature*, Switzerland.

United States Government, Executive Office of the President (2025). Executive Order 14279 of April 23, 2025 Reforming Accreditation To Strengthen Higher Education, 90 FR 17529, 2025-07376 Pages 17529-17532.

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2025). Harnessing the power of AI in qualitative research: Exploring, using and redesigning ChatGPT. *Computers in Human Behavior: Artificial Humans*, 4, 100144. https://doi.org/10.1016/j.chbah.2025.100144

About the Authors

Christopher R. Drue, Associate Director for Teaching Evaluation, Rutgers University, chris.drue@rutgers.edu

Michele Moser Deegan, Associate Vice President of Academic Assessment and Accreditation, Rutgers University, michele.deegan@rutgers.edu

Appendix

Copilot Prompt A (For Copilot Tests 1, 2)
Review the uploaded document and tell me whether it fits into Standard I.1, Standard I.2, Standard I.3, Standard I.4, some combination of these, or none of these standards. Choose between these two options: “Clear Match” or “No Match" (include partial fits in No Match). Summarize the results in a table. [Followed by relevant section of MSCHE Evidence Expectations by Standard Guidelines (Middle States Commission on Higher Education, 2023a)]

Copilot Prompt B (For Copilot Tests 1, 2)
Review the uploaded document and tell me whether it fits into Standard III.1, Standard III.2, Standard III.3, some combination of these, or none of these standards. Choose between these two options: “Clear Match” or “No Match" (include partial fits in No Match). Summarize the results in a table and include a final summary table. [Followed by relevant section of MSCHE Evidence Expectations by Standard Guidelines (Middle States Commission on Higher Education, 2023a)]

Copilot Prompt C (For Copilot Tests 1, 2)
Review the uploaded document and tell me whether it fits into Standard III. 4, Standard III.5, Standard III.6, Standard III.7, Standard III.8, some combination of these, or none of these standards. Choose between these two options: “Clear Match” or “No Match" (include partial fits in No Match). Summarize the results in a table and include a final summary table. [Followed by relevant section of MSCHE Evidence Expectations by Standard Guidelines (Middle States Commission on Higher Education, 2023a)]

Copilot Prompt D (For Copilot Tests 1, 2)
Review the uploaded document and tell me whether it fits into Standard V.1, Standard V.2, Standard V.3, Standard V.4, Standard V.5, some combination of these, or none of these standards. Choose between these two options: “Clear Match” or “No Match" (include partial fits in No Match). Summarize the results in a table. [Followed by relevant section of MSCHE Evidence Expectations by Standard Guidelines (Middle States Commission on Higher Education, 2023a)]

Copilot Prompt E (For Copilot Test 3)
Look at the uploaded file and decide whether it fits into Standard I.1, Standard I.2, Standard I.3, Standard I.4, Standard III.1, Standard III.2, Standard III.3, Standard III.4, Standard III.5, Standard III.6, Standard III.7, Standard III.8, Standard V.1, Standard V.2, Standard V.3, Standard V.4, Standard V.5, some combination of these, or none of these. Choose between: “Clear Match” or “No Match" (include partial fits in No Match). Show the results in a table that lists each sub-standard on the left-most column, “Clear Match” or “No Match" in the middle column, and 1 sentence of explanation for each finding in the right-most column.

Standard I.1 requires institutions to maintain a clearly defined mission and set of goals that are collaboratively developed, reflect both internal and external contexts, and are formally approved by the governing body. These missions and goals must guide institutional decision-making in areas such as

planning, resource allocation, curriculum development, and outcome definition. They should support scholarly and creative activity appropriate to the institution and be widely communicated to internal stakeholders. Supporting evidence includes the mission statement with revision history and committee involvement, documentation of governing body approval, communications to stakeholders, and samples demonstrating alignment between mission, goals, planning, budgeting, and institutional outcomes.

Standard I.2 emphasizes that institutional goals must be realistic, appropriate to higher education, and clearly aligned with the mission. Evidence typically includes the most recent strategic or institutional effectiveness plans, with goals cross-referenced to the mission.

Standard I.3 focuses on goals related to student learning and achievement, including retention, graduation, transfer, and placement rates. These goals should incorporate diversity, equity, and inclusion principles and be supported by administrative and educational programs. Institutions must demonstrate progress using disaggregated student achievement data over time, and show alignment between mission, strategic goals, DEI principles, and resource allocation. Supporting documentation includes IPEDS data, alternative completion measures, standardized exam pass rates, and budget analyses.

Standard I.4 requires periodic assessment of the mission and goals to ensure they remain relevant and achievable. Evidence includes documentation of strategic planning processes and regular evaluations of institutional mission and goals.

Standard III.1 requires that all academic programs—whether certificate, undergraduate, graduate, or professional—are designed to provide a coherent learning experience and promote integration of knowledge. Programs must assign credit hours that reasonably reflect student workload and include sufficient content and length appropriate to credential objectives. Institutions must document policies and procedures for credit hour assignment across all formats and modalities, demonstrate consistent application, and ensure public access to accurate information in compliance with federal regulations. Standard III.2 emphasizes that student learning experiences must be designed, delivered, and assessed by qualified faculty and professionals who are effective in teaching, assessment, and scholarly activity. Institutions must maintain sufficient staffing to ensure program continuity and coherence, support professional development, and conduct regular, equitable evaluations based on clear criteria. Evidence includes faculty qualifications, workload data, student-to-faculty ratios, instructional expenditures, and documentation of program reviews, course evaluations, and faculty training.

Standard III.3 requires that academic programs be clearly described in official publications so students can understand degree requirements and expected time to completion. Institutions must provide accessible catalogs, enrollment data, and trend analyses of student progress, including time to degree and credit accumulation. Additional educational offerings such as dual enrollment, career and technical education, and ESL programs must also be documented and analyzed by relevant populations.

Standard III.4 mandates that institutions offer sufficient learning experiences and resources to support academic programs and student progress. This includes advising materials, syllabi, and comprehensive

library and learning resources across all instructional sites and formats. Institutions must assess the effectiveness of resource policies, provide bibliographic instruction and information literacy programs, and maintain agreements for shared resources. For distance education, institutions must ensure qualified faculty, student identity verification, instructional quality, access to support services, and regular assessment of design and technology infrastructure, with related expenditures tracked over time.

Standard III.5 requires institutions offering undergraduate education to provide a general education program—either standalone or integrated—that broadens students' intellectual horizons, enhances cultural and global awareness, and prepares them for sound judgment beyond their academic discipline. The curriculum must ensure students acquire core competencies in communication, scientific and quantitative reasoning, critical thinking, technological literacy, and information literacy. It should also include study of values, ethics, and diverse perspectives, aligned with the institution's mission. Non-U.S. institutions without formal general education must show that students still demonstrate these competencies. Supporting documentation includes curriculum maps, catalog descriptions, learning outcomes evaluations, program revision history, and records of newly offered courses.

Standard III.6 applies to institutions offering graduate professional education and requires opportunities for students to engage in research, scholarship, and independent thinking, guided by appropriately credentialed faculty. Evidence includes graduate-level learning outcomes, policies on research and assistantships, faculty qualifications, and expenditure data related to research.

Standard III.7 mandates that institutions maintain appropriate oversight of student learning experiences provided by third-party entities. Institutions must have clear policies for approving such providers, maintain records of current partnerships and contracts, disclose third-party involvement publicly, and regularly evaluate the quality and effectiveness of these learning experiences.

Standard III.8 requires periodic assessment of student learning experiences across all student populations. Institutions must demonstrate regular evaluation, analysis of results, and follow-up actions to ensure continuous improvement.

Standard V.1 requires institutions to clearly articulate student learning outcomes at both the institutional and program levels. These outcomes must be interconnected with one another, aligned with educational experiences, and consistent with the institution's mission. Supporting evidence includes curriculum maps and matrices showing how outcomes relate across programs and general education.

Standard V.2 emphasizes the need for systematic, faculty-led assessment processes that evaluate student achievement of learning goals. Institutions must define appropriate outcomes, establish defensible assessment standards, and demonstrate how their programs prepare students for careers, meaningful lives, and further education. They must collect and share assessment data with stakeholders and sustain ongoing evaluation efforts.

Standard V.3 calls for the use of disaggregated assessment results to improve learning outcomes, student achievement, and overall educational effectiveness. Institutions should analyze data across all programs with sufficient enrollment, apply varied assessment approaches, and interpret trends over time, including post-completion outcomes such as job placement rates.

Standard V.4 applies when third-party providers are involved in assessment services. Institutions must ensure proper review and approval processes, maintain formal agreements, and periodically evaluate the effectiveness of these external services.

Standard V.5 requires regular review of the institution's own assessment policies and procedures to ensure they contribute meaningfully to educational improvement. Institutions must document these evaluations, consider findings, and take appropriate follow-up actions.

Vendor AI Prompt A (For Vendor AI Test 1)
Look only at the file titled "FILE NAME" and tell me whether it fits into Standard I.1, Standard I.2, Standard I.3, Standard I.4, Standard III.1, Standard III.2, Standard III.3, Standard III.4, Standard III.5, Standard III.6, Standard III.7, Standard III.8, Standard V.1, Standard V.2, Standard V.3, Standard V.4, Standard V.5 some combination of these, or none of these standards. Choose between these two options: "Clear Match" or "No Match" (include partial fits in No Match). Show the results in a table that lists each sub-standard on the left-most column, "Clear Match" or "No Match" in the middle column, and 1 sentence of explanation for each finding in the right-most column. [Followed by entire text of MSCHE Evidence Expectations by Standard Guidelines (Middle States Commission on Higher Education, 2023a)]

Vendor AI Prompt B (For Vendor AI Test 2)
Look only at the file titled "FILE NAME" and decide whether it fits into Standard I.1, Standard I.2, Standard I.3, Standard I.4, Standard III.1, Standard III.2, Standard III.3, Standard III.4, Standard III.5, Standard III.6, Standard III.7, Standard III.8, Standard V.1, Standard V.2, Standard V.3, Standard V.4, Standard V.5, some combination of these, or none of these. Choose between: "Clear Match" or "No Match" (include partial fits in No Match). Show the results in a table that lists each sub-standard on the left-most column, "Clear Match" or "No Match" in the middle column, and 1 sentence of explanation for each finding in the right-most column.

Standard I.1: The mission and goals of an institution are fundamental to its identity and function. They are crafted through collaborative efforts involving all stakeholders responsible for institutional development and improvement. These goals address both internal and external contexts and constituencies and receive approval and support from the governing body. They guide faculty, administration, staff, and governing structures in decision-making related to planning, resource allocation, program and curricular development, and defining institutional and educational outcomes. The mission also supports scholarly inquiry and creative activity at levels appropriate to the institution. It is publicized and widely known among the institution's internal stakeholders. For example, a mission statement's last revision date, the list of mission review committee members, and evidence of their involvement in mission review and revision are crucial. Additionally, evidence of participation and

approval by the governing body, sample communications or publications of the mission statement, and evidence of alignment between mission elements and institutional goals are essential.

Standard I.2: Institutional goals must be realistic, appropriate to higher education, and consistent with the mission. These goals are documented in the most recent institutional strategic plan or institutional effectiveness plan, or other strategic planning or goal-setting documentation. The date of the last update and goals with evidence of their relationship to the mission, such as a crosswalk, are examples of how these goals are structured and aligned with the institution's mission.

Standard I.3: Goals focusing on student learning outcomes and achievement include retention, graduation, transfer, and placement rates. They consider diversity, equity, and inclusion principles and are supported by administrative, educational, and student support programs and services, prioritizing institutional improvement. Evidence of set goals addressing student learning outcomes and attainment measured using student achievement data, such as graduation and retention rates, is vital. Trend analysis of the institution’s progress at meeting established student achievement goals using at least four years of data, disaggregated by relevant populations, is also crucial. Additionally, evidence of alignment between mission, strategic goals, and diversity, equity, and inclusion principles, along with budgetary support and resource allocation, supports these goals.

Standard I.4: Periodic assessment of mission and goals ensures they remain relevant and achievable. Evidence of strategic plan and mission development processes, along with regular evaluation of the mission statement and institutional goals, demonstrates the institution's commitment to continuous improvement and alignment with its mission.

Standard III.1: Academic programs leading to a degree or other recognized higher education credential are designed to foster a coherent student learning experience and promote synthesis of learning. They are assigned a reasonably approximate number of credit hours for the amount of work completed by a student and include sufficient course content and program length appropriate to the objectives of the degree or other credential. Policies, procedures, and methodologies employed for assignment of credit hours, course or program review procedures, and documentation of evaluation processes are examples of how these programs are structured.

Standard III.2: Student learning experiences are designed, delivered, and assessed by faculty and other appropriate professionals who are rigorous and effective in teaching, assessment of student learning, scholarly inquiry, and service. They are qualified for their positions and work, sufficient in number, and provided with opportunities, resources, and support for professional growth and innovation. Regular and equitable reviews based on clear criteria, expectations, policies, and procedures ensure faculty competence and effectiveness.

Standard III.3: Academic programs of study are clearly and accurately described in official publications, enabling students to understand and follow degree and program requirements and expected time to completion. Academic catalogs, trend analysis of academic progress, and descriptions of educational offerings provide clarity and transparency for students.

Standard III.4: Sufficient learning experiences and resources support the institution's programs of study and the academic progress of all student populations. Advising or degree program sheets, syllabi, library/learning resources, and policies for access to information resources ensure students have the necessary support for their academic journey.

Standard III.5: A general education program offers a sufficient scope to draw students into new areas of intellectual experience, expanding their cultural and global awareness and cultural sensitivity. It prepares them to make well-reasoned judgments outside and within their academic field. Documentation of curriculum maps, evaluation of student learning outcomes, and descriptions of educational experiences ensure the program's coherence and effectiveness.

Standard III.6: Graduate professional education provides opportunities for the development of research, scholarship, and independent thinking. Faculty and professionals with appropriate credentials support these opportunities. Graduate-level student learning outcomes, policies related to independent research, and faculty qualifications demonstrate the institution's commitment to advanced education.

Standard III.7: Adequate and appropriate institutional review and approval are required for any student learning opportunities designed, delivered, or assessed by third-party providers. Policies, procedures, and documentation of agreements with third-party providers ensure the quality and integrity of these educational experiences.

Standard III.8: Periodic assessment of the effectiveness of student learning experiences for all student populations is essential. Evidence of regular evaluation and assessment, consideration of results, and follow-up ensures continuous improvement and alignment with institutional goals.

Standard V.1: Clearly stated student learning outcomes at the institution and degree/program levels are interrelated with one another, relevant educational experiences, and the institution's mission. Documentation of student learning outcomes, curriculum maps, and their relationship to each other ensures clarity and coherence in educational objectives.

Standard V.2: Organized and systematic assessments evaluate the extent of student achievement of institutional and degree/program goals. Institutions define appropriate student learning outcomes, collect data on goal achievement, and communicate assessment results to stakeholders. Documentation of assessment processes and communication of results demonstrate the institution's commitment to transparency and accountability.

Standard V.3: Disaggregated assessment results for all student populations are used to improve student learning outcomes, achievement, and educational effectiveness. Analysis of assessment results, assessment approaches, and disaggregated data help interpret educational effectiveness and guide improvements.

Standard V.4: If applicable, adequate and appropriate institutional review and approval of assessment services designed, delivered, or assessed by third-party providers are necessary. Policies, procedures,

agreements, and evidence of periodic assessment ensure the quality and reliability of third-party assessment services.

Standard V.5: Periodic assessment of the effectiveness of assessment policies and processes is crucial for improving educational effectiveness. Evidence of periodic assessment, consideration of results, and follow-up ensures that assessment practices remain relevant and effective in supporting institutional goals.