2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC

Host Sponsor UNC Charlotte

Paper ID #52533

Exploring AI-Powered Recognition of Handwritten Reasoning and Reflection for Authentic Assessment to Sustain the Pedagogical Effectiveness of Writing-to-Learn in the Generative AI Era

Prof. Wei Zheng, Jackson State University

Dr. Wei Zheng is a Professor of Civil Engineering at Jackson State University. He received his Ph.D. in Civil Engineering from the University of Wisconsin–Madison in 2001 and has more than ten years of industrial experience before joining academia in 2005. His research mainly focuses on applying probabilistic machine learning to decision-making under uncertainty. His work integrates expertise in civil engineering, AI model development, and educational data analytics to improve engineering decision systems and personalized learning support. His current research in civil engineering focuses on using generative AI technologies to extract knowledge from historical engineering reports and records to support data-driven decision-making. He is also leading two ongoing NSF-funded projects and conducts research in two key areas: (1) Engaging students in open-ended responses and reflective writing to strengthen higher-order thinking and learning skills, and (2) Developing AI-powered automatic assessment tools to evaluate these responses, with the ultimate goal of providing adaptive support with bias mitigation and cultural relevance enhancement through AI-powered education to promote equitable and effective learning.

Hrijan Pandit

Exploring AI-Powered Recognition of Handwritten Reasoning and Reflection for Authentic Assessment to Sustain the Pedagogical Effectiveness of Writing-to-Learn in the Generative AI Era

Wei Zheng[1] and Hrijan Pandit[2]

[1] Department of Civil and Environmental Engineering, Jackson State University, Jackson, MS, USA

[2] Department of Computer and Information Science, University of Mississippi, 201 Weir Hall, University, MS 38677, USA

Abstract

The increasing availability of generative artificial intelligence has raised concerns about the authenticity of typed student writing, particularly for Writing-to-Learn (WTL) activities intended to capture students' learning processes rather than polished final answers. Handwritten responses provide a potential means of preserving authenticity, but their instructional use at scale is limited by the difficulty of digitization and analysis. This paper investigates the feasibility of using an AI-based handwriting-to-text transcription pipeline to support handwritten WTL artifacts in undergraduate engineering education. The study is situated within a multi-cycle WTL pedagogy and focuses on three reflective learning processes: learning plans, learning evaluations, and learning reflections. Handwritten responses were collected from junior- and senior-level engineering courses and digitized using a cloud-based vision–language transcription pipeline with human-in-the-loop validation. Evaluation was conducted at the student-submission level. Results show that the proposed pipeline consistently produces accurate and semantically faithful transcripts of handwritten reflective writing, outperforming baseline OCR-based approaches under authentic classroom conditions. These findings suggest that AI-assisted handwriting transcription can enable scalable use of handwritten Writing-to-Learn activities while preserving pedagogical intent and authenticity. The work contributes evidence that handwriting, supported by appropriate AI tools, remains a viable response modality for reflective learning in the generative-AI era. In addition, by evaluating the system on authentic HBCU classroom data, this study helps address the underrepresentation of such educational contexts in handwriting-recognition research and provides initial insight into the robustness of AI-based transcription in diverse educational settings.

1. Introduction and Literature Review

The rapid advancement of generative artificial intelligence (AI) has fundamentally changed how students produce written responses in higher education. In engineering education, open-ended writing activities—long used to support Writing-to-Learn (WTL) pedagogy—are increasingly vulnerable to AI-generated or AI-assisted text, raising concerns about the validity of student writing as evidence of learning processes rather than polished linguistic output. These concerns also have an equity dimension, as differences in access to generative AI tools and variation in how students engage with them may amplify disparities in learning processes. This challenge is particularly acute for instructional activities designed to elicit metacognitive and motivational processes, such as learning plans, post-assessment evaluations, and reflective writing, where authenticity of student thinking is central to pedagogical intent [1], [2].

One instructional response to this challenge has been renewed interest in handwritten student responses. Handwriting introduces cognitive, temporal, and motor constraints that make real-time AI text generation impractical and encourage students to externalize their own reasoning, planning, and reflection processes. Prior educational research has long suggested that handwriting can support deeper cognitive engagement and reflection compared to typing, particularly for learning activities that emphasize process rather than final answers [3]. However, while handwriting may help restore authenticity, it introduces a practical barrier: handwritten artifacts are difficult to analyze, archive, and scale for instructional use without automated transcription.

Recent advances in handwriting recognition and vision–language modeling have made it technically feasible to convert handwritten text into machine-readable form. A growing body of work has explored handwritten text recognition (HTR) using convolutional and transformer-based architectures, demonstrating strong performance on benchmark datasets and controlled evaluation settings [4], [5]. In educational contexts, several studies have proposed pipelines that combine handwriting recognition with automated grading or feedback generation, often integrating optical character recognition (OCR) or vision–language models with large language models [6], [7]. These systems show promise for automating assessment of handwritten student work, particularly for short answers or structured problem-solving tasks.

Despite these advances, existing work exhibits several important limitations when viewed through the lens of Writing-to-Learn pedagogy. First, most prior studies focus on correctness-oriented tasks, such as mathematical derivations, short explanations, or exam-style responses, rather than long-form reflective writing. Second, evaluation metrics are predominantly technical—character error rate, word error rate, or grading agreement—without examining whether the *meaning and intent* of students' writing, especially metacognitive and motivational content, are preserved after transcription [5], [8]. Third, many evaluations rely on benchmark datasets or carefully controlled data collection, leaving open questions about robustness to authentic classroom handwriting, including variability in handwriting style, organization, and scan quality.

A further gap concerns student population and context. Handwriting recognition systems are rarely evaluated using data from underrepresented student populations, including students at Historically Black Colleges and Universities (HBCUs). Given well-documented disparities in AI system performance across populations, the absence of evaluation using authentic HBCU classroom data raises concerns about generalizability and equity in educational AI deployment [2], [9]. In this study, the use of HBCU classroom data is intentional, enabling an initial examination of whether handwriting transcription models remain robust in diverse and underrepresented educational contexts.

Taken together, the existing literature demonstrates that while handwriting recognition technology has matured, its educational validity for Writing-to-Learn artifacts remains largely unexamined. In particular, there is little empirical evidence addressing whether AI-based handwriting transcription can preserve the semantic, metacognitive, and motivational content of learning plans, learning evaluations, and learning reflections—artifacts explicitly designed to capture students' learning processes rather than their final answers.

This study addresses that gap by evaluating an AI-based handwriting-to-text pipeline within an authentic Writing-to-Learn pedagogical framework implemented in undergraduate engineering courses. Rather than treating handwriting recognition as a standalone technical problem, the study examines transcription performance in terms of both technical accuracy and pedagogical usefulness, using real classroom data. The resulting evidence directly informs whether handwritten Writing-to-Learn activities can be feasibly and responsibly integrated into AI-supported instructional workflows in the generative-AI era.

2. Writing-to-Learn Pedagogical Framework and Instructional Context

Building on the literature and gaps identified in Section 1, this study situates handwriting recognition within a Writing-to-Learn (WTL) pedagogical framework designed to foreground students' learning processes rather than final performance outcomes. While prior work on handwriting recognition has largely emphasized transcription accuracy or automated grading, the present work adopts a pedagogical lens, focusing on how handwritten artifacts function as evidence of planning, self-monitoring, and reflection in the learning process.

Writing-to-Learn pedagogy is grounded in the premise that structured writing activities can promote deeper conceptual understanding by encouraging learners to externalize their thinking, identify misconceptions, and regulate their learning strategies. In engineering education, WTL has been used to support students' metacognitive development through activities such as learning plans, post-assessment evaluations, and reflective writing, which emphasize *how* students learn rather than simply *what* they produce. These activities are particularly valuable in the generative-AI era, where polished written responses may no longer reliably reflect students' own reasoning processes [1], [2].

In the instructional design examined in this study, WTL is implemented as a multi-cycle learning process that unfolds across the semester. Each cycle consists of five instructional components: (1) homework or practice problems involving open-ended domain-specific questions, (2) a learning plan in which students articulate how they will prepare for an upcoming quiz, (3) a quiz that includes both traditional problems and open-ended responses, (4) a learning evaluation completed after the quiz is graded and returned, and (5) a learning reflection focused on students' learning strategies and processes. Together, these components form an integrated pedagogical sequence that supports planning, performance, evaluation, and reflection.

The present paper focuses specifically on three handwritten Writing-to-Learn artifacts produced within this cycle: the learning plan, the learning evaluation, and the learning reflection. These artifacts were selected because they are explicitly designed to elicit metacognitive and motivational information, including students' goal setting, self-diagnosis of errors, strategy selection, and time management decisions. Unlike homework or quiz responses, which emphasize content correctness, these artifacts capture students' internal learning processes and are therefore particularly sensitive to issues of authenticity and meaning preservation.

Handwriting was intentionally introduced as the response modality for these artifacts to mitigate the influence of generative AI during text production. As discussed in Section 1, handwriting imposes cognitive and procedural constraints that make real-time AI generation impractical and encourage students to engage directly with their own thinking processes [3]. However, requiring handwritten responses creates practical challenges for instructors, including increased workload for reading, archiving, and analyzing student writing. This tension motivates the integration of AI-based handwriting-to-text transcription as a supporting layer rather than a replacement for pedagogical judgment.

Within this framework, handwriting recognition is not treated as an end in itself, but as an enabling technology that may allow handwritten Writing-to-Learn artifacts to be incorporated into scalable instructional workflows. The central question is whether AI-generated transcripts preserve sufficient semantic and pedagogical fidelity to support analysis of students' learning processes. Addressing this question requires evaluating handwriting recognition not only in terms of technical accuracy, but also in terms of its alignment with the goals of Writing-to-Learn pedagogy.

In addition, handwritten Writing-to-Learn activities may support diverse learning styles by encouraging slower, more deliberate processing and reducing reliance on external tools such as generative AI. This modality can help create a more equitable assessment environment, particularly for students who may have unequal access to AI tools or who benefit from process-oriented forms of expression. As a result, handwriting not only supports authenticity but also reinforces inclusive participation in reflective learning practices.

3. Research Questions

Building on the instructional motivation and pedagogical context established in Sections 1 and 2, this study evaluates whether an AI-based handwriting-to-text pipeline can support handwritten Writing-to-Learn artifacts in ways that preserve technical accuracy, semantic fidelity, and instructional usefulness. The pipeline is treated as an enabling tool for analyzing handwritten WTL artifacts, not as an autonomous grading system.

Accordingly, this study is guided by the following research questions:
RQ1. *To what extent can an AI-based handwriting-to-text pipeline accurately transcribe authentic, long-form handwritten Writing-to-Learn artifacts produced by undergraduate engineering students?*

This question addresses technical feasibility under realistic classroom conditions, including variability in handwriting style, organization, and scan quality, rather than controlled benchmark performance.

RQ2. *To what extent do AI-generated transcripts preserve the semantic meaning of students' handwritten learning plans, learning evaluations, and learning reflections?*

This question focuses on pedagogical fidelity, emphasizing whether the meaning and intent of metacognitive and motivational writing are retained after transcription, beyond surface-level character or word accuracy.

RQ3. *How do students perceive the fidelity of AI-generated transcripts relative to their original handwritten responses?*

This question incorporates student-centered validation to examine whether students recognize their own thinking, intentions, and reflections in the transcribed text, providing an additional lens on educational validity.

RQ4. *Is the performance of the handwriting-to-text pipeline sufficiently robust to support scalable use of handwritten Writing-to-Learn artifacts within an instructional workflow?*

Rather than evaluating downstream automated assessment, this question examines whether transcription quality is adequate to enable instructor review, archival, and future integration with rubric-based or human-in-the-loop evaluation.
Together, these research questions align with the pedagogical goals of Writing-to-Learn and directly address the knowledge gaps identified in prior work. They emphasize meaning preservation, instructional usability, and student experience, rather than algorithmic novelty or isolated accuracy metrics. This framing ensures that the evaluation of the handwriting-to-text pipeline remains grounded in educational purpose and instructional decision-making.

While not formulated as a separate research question, the study implicitly examines robustness in an HBCU instructional context, providing initial insight into the performance of handwriting transcription systems in underrepresented educational settings.

4. Data Sources, Participants, and Study Context

This study draws on authentic classroom data collected as part of a Writing-to-Learn (WTL) instructional implementation in undergraduate engineering courses. The instructional design, data collection, and analysis procedures were embedded within regular course activities and aligned with the pedagogical framework described in Section 2. The purpose of this section is to describe the study context, participants, data sources, and scope of analysis, rather than to introduce methodological details of the transcription pipeline itself.

4.1 Course Context and Participants

Data were collected from two undergraduate engineering courses offered during the same academic term. One course was a **junior-level course**, and the other was a **senior-level course**, both of which incorporated Writing-to-Learn activities as part of normal instructional practice. Across the two courses, approximately **forty students** participated, with each course contributing roughly twenty students. These courses were offered at a Historically Black College and University (HBCU), and the student population reflects the institutional context in which the pedagogy was implemented.

The handwritten responses for the learning plan, learning evaluation, and learning reflection were produced as part of required course assignments implementing the Writing-to-Learn pedagogy, which was reviewed and approved by the Institutional Review Board (IRB). The use of these course artifacts for research purposes was voluntary. Students were informed through a consent form that they could choose whether to permit their coursework to be used for research, that declining permission would not affect their course grades or standing, and that all data would be de-identified prior to analysis. Only coursework from students who voluntarily granted permission for research use through the consent process was included in this study.

Eight of the fourteen enrolled students in the junior-level course and sixteen of the twenty-one enrolled students in the senior-level course submitted handwritten responses for the learning plan, learning evaluation, and learning reflection as the final Writing-to-Learn assignment in their respective courses. These handwritten submissions constitute the dataset used for the analysis reported in this paper.

4.2 Writing-to-Learn Artifacts and Handwritten Data

The data analyzed in this paper consist of handwritten Writing-to-Learn artifacts produced by students as part of a multi-cycle WTL pedagogy. As discussed earlier, each WTL cycle includes multiple instructional components; however, the present study focuses specifically on three handwritten artifacts that are explicitly designed to capture students' learning processes:

1. Learning Plans, in which students describe how they intend to prepare for an upcoming quiz, identify concepts requiring further study, and articulate strategies and time management plans.
2. Learning Evaluations, completed after a quiz is graded and returned, in which students analyze their performance, identify errors and misconceptions, and reflect on why difficulties occurred.
3. Learning Reflections, in which students reflect on the effectiveness of their learning strategies, evaluate their effort and engagement, and consider adjustments for future learning.

These three artifacts were selected because they emphasize metacognitive and motivational processes, rather than content correctness, and therefore place particularly strong demands on semantic fidelity during transcription.

Although the broader WTL pedagogy also includes open-ended responses embedded in homework and quizzes, those artifacts are outside the scope of the present paper. The focus here is intentionally limited to handwritten responses associated with learning plans, evaluations, and reflections.

4.3 Handwriting Collection and Digitization

Students completed the selected Writing-to-Learn activities by hand using paper-based formats. Each student produced one handwritten response for each artifact within the WTL cycle examined in this study. The handwritten pages were subsequently scanned into PDF format using standard document scanners, preserving the natural variability in handwriting style, layout, and organization typical of classroom work.

No constraints were imposed on handwriting style, spacing, or formatting beyond those normally present in the course. As a result, the dataset reflects realistic classroom conditions, including variation in legibility, structure, and annotation practices.

4.4 Scope of the Present Study

The handwritten PDF files served as the input to the handwriting-to-text transcription pipeline described in Section 5. The outputs of that pipeline—AI-generated transcripts—were then used for evaluation and analysis. Importantly, the transcription pipeline is treated in this study as a supporting tool that enables analysis of handwritten WTL artifacts; it is not used to grade students or to replace instructor judgment.

The scope of the present paper is limited to handwritten textual content. Although students in later instructional cycles were given the option to provide oral responses, and although some engineering tasks involve hand-drawn diagrams or symbolic representations, those data are intentionally excluded here and will be addressed in separate work.

By clearly defining the study context and data sources, this section establishes the boundary conditions under which the results in subsequent sections should be interpreted. The use of data from an HBCU instructional context is intentional and aligns with the study's focus on evaluating transcription robustness under realistic and diverse classroom conditions. Because handwriting styles, organization, and expression can vary substantially across student populations, incorporating data from an underrepresented educational setting provides a more comprehensive basis for assessing the generalizability and reliability of AI-based handwriting transcription systems.

5. Handwriting-to-Text Transcription Pipeline

This section describes the handwriting-to-text transcription pipeline used in this study. Consistent with the instructional goals outlined in Sections 1–4, the pipeline is designed as an **enabling tool** that supports the use of handwritten Writing-to-Learn (WTL) artifacts in scalable instructional workflows. The system is not intended to replace instructor judgment or to perform automated grading; instead, it focuses on reliable transcription, identity-aware text extraction, and human-validated outputs suitable for pedagogical analysis.

5.1 System Overview and Design Rationale

The proposed pipeline is a cloud-deployed handwriting recognition system designed to convert students' handwritten, open-ended responses into structured digital text for downstream evaluation (see Fig.1) . It is motivated by instructional settings in which students submit handwritten responses via scanned documents or image-based platforms, such as Google Forms. In these settings, instructors often receive multiple handwritten submissions bundled into a single document, creating practical challenges for reading, indexing, and analysis.

The pipeline enables instructors to upload a single multi-page document containing multiple students' handwritten responses, each labeled with a visible student identity code, and automatically obtain text outputs indexed by student identifiers. A human-in-the-loop review step is incorporated to allow instructors to validate and correct extracted text prior to any instructional use. This design reflects three practical constraints common in educational settings: reliance on publicly available technologies without retraining domain-specific models, minimal disruption to existing instructional workflows, and mandatory human oversight to preserve assessment validity and instructional trust [1]. At its current stage, the system focuses on robust handwriting recognition and identity-aware text extraction, which together constitute the core methodological contribution of this paper.

To contextualize the performance of the proposed system, the pipeline is evaluated alongside three baseline configurations, including traditional OCR, OCR with language-model post-processing, and a naive vision–language model baseline without structural constraints.

The pipeline is designed to operate at the student-submission level, rather than at the individual page level, reflecting the structure of Writing-to-Learn artifacts and enabling end-to-end reconstruction of multi-part student responses.

5.2 Input Format and Data Assembly

Student responses are collected as image files (e.g., JPEG or PNG) through online platforms such as Google Forms, where each submission contains handwritten text along with a visible student identity code written directly on the page. Image-based submission of handwritten work has become increasingly common in engineering education, particularly when instructors seek to preserve authentic student reasoning while enabling digital collection.

For pipeline processing, individual image submissions are programmatically assembled into a single multi-page PDF document, with each page corresponding to one student response. PDF-based batch processing is widely used in document understanding and handwriting recognition workflows because it supports scalable inference while preserving page-level structure and metadata [7]. The current implementation assumes legible handwritten text and clearly separated student identity codes. Handling diagrams, mathematical notation, or highly stylized cursive writing is beyond the scope of the present study and is discussed as future work.

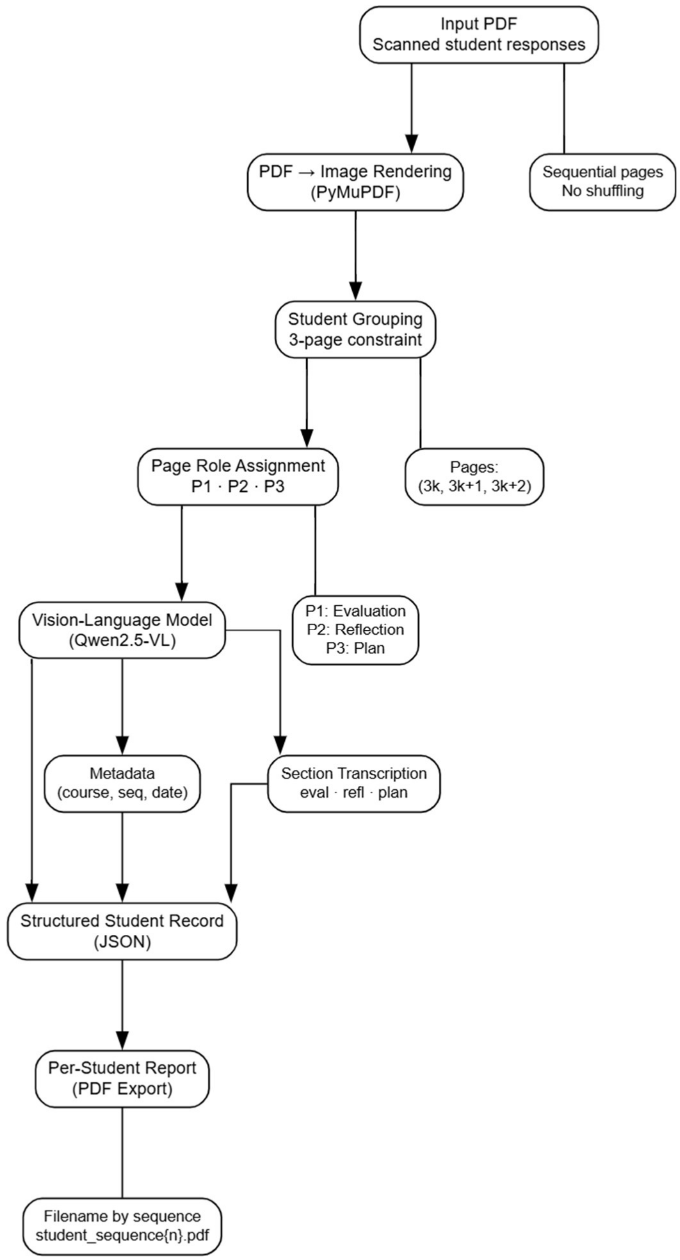


Figure 1: Overview of the proposed training-free vision-language pipeline for structured transcription of handwritten student submissions.

5.3 Cloud-Hosted Vision–Language Model Backend

The handwriting recognition component of the pipeline is powered by a large vision–language model (VLM), specifically Qwen2.5-VL, deployed on Amazon Web Services (AWS). Vision–language models integrate visual perception and natural language understanding within a unified architecture and have demonstrated strong performance on document understanding tasks involving unconstrained layouts and handwritten content [6], [10].

Deploying the model in a cloud environment enables scalable inference, consistent execution environments, and flexible access for instructional use, which aligns with common deployment practices for large AI models in applied educational settings [11]. Rather than relying solely on traditional optical character recognition (OCR) techniques, the VLM is prompted to jointly interpret visual handwriting and semantic context. This design choice is motivated by the variability in student handwriting styles and page layouts typically observed in authentic classroom submissions, which often challenge character-based OCR systems [7].

5.4 Prompt-Based Identity-Aware Text Extraction

A key methodological feature of the pipeline is the use of structured prompt engineering to support identity-aware text extraction. The VLM is prompted to simultaneously identify (1) the student identity code visible in the handwritten response and (2) the corresponding handwritten textual content. The model is instructed to return outputs in a structured format, such as JSON-like key–value pairs, explicitly linking each transcribed response to its student identifier.

Prompt-based control of large language and vision–language models has been shown to improve output consistency and task alignment for information extraction tasks [12]. In this study, structured prompting is used not as a modeling innovation, but as a practical mechanism to ensure reliable student attribution when processing multi-student handwritten documents. This approach avoids post-hoc heuristic matching and enables direct indexing of responses during inference.

The proposed structured pipeline uses the same underlying Qwen2.5-VL model as the naive baseline; performance differences arise from the introduction of structured prompting, identity-aware extraction, and document-level grouping rather than changes in model architecture. In addition, the pipeline enforces multi-page student-level grouping, ensuring that consecutive pages corresponding to a single student submission are processed and reconstructed as a unified record. The pipeline explicitly separates metadata (e.g., course identifier, sequence number, date) from response content, preventing leakage of header information into student writing during transcription. These structural design choices—identity-aware prompting, metadata separation, and student-level grouping—constitute the primary contribution of the proposed pipeline.

5.5 Output Representation

The pipeline produces structured text files indexed by student identifiers. Each record contains the student ID, the transcribed response text, and optional metadata such as page number or image index. This standardized representation enables direct integration with downstream instructional workflows, including manual review, rubric-based scoring, and natural language processing–based evaluators. Importantly, the output format is designed to remain transparent and editable, supporting instructor oversight at all stages.

It should be noted that the baseline methods described in this study—including traditional OCR, OCR with language-model post-processing, and naive vision–language model prompting—are included solely for evaluation purposes and are not implemented as deployable instructional systems. These configurations lack student-level grouping, identity-aware extraction, and structured outputs required for instructional use. In contrast, the proposed structured VLM pipeline is designed as an end-to-end system that supports instructor workflows and enables practical deployment and dissemination in Writing-to-Learn contexts.

5.6 Human-in-the-Loop Validation

Consistent with best practices in educational assessment, the pipeline incorporates a human-in-the-loop validation step. Instructors review extracted text, correct transcription errors, and confirm student attribution before using the results for grading or analysis. Human oversight is widely recommended in AI-assisted educational systems to ensure transparency, fairness, and pedagogical appropriateness, particularly in high-stakes assessment contexts [1]. In this study, human validation is treated as an integral component of the methodology rather than an optional post-processing step.

5.7 Current Scope and Ongoing Extensions

The present implementation establishes a validated handwriting recognition and identity-aware extraction pipeline that enables the conversion of handwritten Writing-to-Learn responses into machine-readable text. Ongoing work includes refinement of the instructor interface, expanded documentation, and tighter integration with automated assessment pipelines. These extensions are beyond the scope of the present paper and are mentioned to situate the methodology within a broader instructional deployment context.

6. Evaluation Design and Metrics

The purpose of the evaluation in this study is to determine whether the proposed handwriting-to-text transcription pipeline is sufficiently reliable and semantically faithful to support Writing-to-Learn (WTL) pedagogy in authentic classroom settings. Consistent with the research questions outlined in Section 3, the evaluation is framed around end-to-end instructional usability rather than isolated character-level recognition performance. Accordingly, the evaluation design emphasizes student-level correctness, semantic fidelity, and robustness under realistic instructional conditions. Because the evaluation uses authentic handwritten submissions from an HBCU instructional context, the analysis also provides an initial test of pipeline robustness in a student population underrepresented in handwriting-recognition research.

6.1 Evaluation Setup

The evaluation treats the proposed system as a training-free vision–language document transcription and structuring pipeline operating on authentic handwritten student submissions. Each student submission consists of three consecutive handwritten pages corresponding to a Learning Evaluation, Learning Reflection, and Learning Plan, as defined in the Writing-to-Learn framework described earlier. These pages are scanned and combined into a single PDF file, with submissions ordered sequentially by student.

The pipeline performs three core functions: (1) transcription of handwritten content, (2) extraction of document-level metadata, such as course identifier, sequence information, and date when present, and (3) reconstruction of structured student-level outputs by grouping every three consecutive pages into a single submission. Importantly, no task-specific fine-tuning or domain-specific training data are used. This design reflects the intended instructional deployment context, in which instructors rely on general-purpose AI tools and practical structural workflows rather than custom model training.

To contextualize performance, we compare four processing configurations. The first three serve exclusively as evaluation baselines, whereas the fourth is the operational system designed for instructional use. The compared methods are traditional optical character recognition (OCR) using Tesseract, Tesseract with Qwen2.5-VL post-processing, a naive Qwen2.5-VL baseline using page-wise prompting without structural constraints, and the proposed structured VLM pipeline. Because the naive baseline and the proposed pipeline use the same Qwen2.5-VL backbone, performance differences between them reflect the contribution of structural design choices—such as identity-aware prompting, metadata separation, and student-level grouping—rather than differences in model capability.

6.2 Evaluation Protocol

Evaluation was conducted through manual verification of randomly sampled student submissions. Rather than evaluating all collected data, the study reports results on subsets of 5, 10, and 20 student submissions, reflecting increasing evaluation scale while maintaining feasibility for careful human inspection. This sampling strategy is consistent with prior work emphasizing human-verified evaluation for educational AI systems, where correctness must remain interpretable and pedagogically meaningful rather than statistically opaque [1], [2].

A student submission is considered correctly processed only if all of the following conditions are satisfied:

1. Metadata fields, such as course name, sequence number, and date, are correctly extracted from the first page or correctly marked as N/A if absent;
2. The Learning Evaluation, Learning Reflection, and Learning Plan texts are correctly associated with their respective handwritten pages;
3. The transcribed content is semantically faithful to the handwritten response, allowing for minor spelling or grammatical errors that do not alter meaning; and
4. No header leakage occurs, such as metadata being repeated inside response bodies.

This protocol intentionally emphasizes end-to-end correctness at the student-submission level rather than isolated page-level transcription accuracy, because the instructional utility of a transcript depends not only on character-level fidelity but also on correct attribution, ordering, and association of text with the intended artifact type. A transcript that correctly recovers content but assigns it to the wrong student or the wrong section is not usable for instructional purposes. Accordingly, each submission must remain interpretable as a complete and coherent unit.

Human reference transcripts were created by a research assistant following consistent transcription guidelines. The transcription procedure required content to be transcribed verbatim where legible, ambiguous characters to be explicitly flagged, and the three-section structure—Learning Evaluation, Learning Reflection, and Learning Plan—to be preserved. No automated preprocessing was applied to the reference transcripts prior to evaluation. Overall accuracy is computed as the proportion of student submissions that satisfy all correct conditions:

$$\text{Accuracy} = \frac{\text{Number of correctly processed submissions}}{\text{Total evaluated submissions}}$$

6.3 Evaluation Metrics

The primary evaluation metric is structured transcription accuracy at the student level, reflecting whether a complete multi-page submission is correctly transcribed, structured, and attributed. This metric directly corresponds to the instructional use case: instructors must be able to read, interpret, and analyze student responses without manually reconstructing document structure or correcting systemic errors.

While traditional metrics such as character error rate (CER) or word error rate (WER) are commonly reported in handwriting recognition studies, they are insufficient for Writing-to-Learn artifacts that emphasize metacognitive and motivational meaning rather than surface correctness. A transcript with low CER may still be pedagogically unusable if metadata is misassigned, pages are mixed across students, or section boundaries are corrupted. For this reason, the evaluation prioritizes semantic fidelity and structural correctness over isolated character-level accuracy.

6.4 Rationale for Baseline Comparisons

The selected baselines reflect common approaches instructors or developers might adopt when attempting to process handwritten student work. Traditional OCR represents a widely accessible but text-centric solution that lacks document-level understanding. Tesseract with Qwen2.5-VL post-processing reflects a hybrid approach increasingly used in practice, where a language model is employed to repair noisy OCR output. The naive Qwen2.5-VL baseline captures the capability of a modern vision–language model when structural constraints are not enforced. In contrast, the proposed structured VLM pipeline is the only method designed as an operational instructional system, incorporating identity-aware prompting, metadata separation, multi-page student-level grouping, and human-in-the-loop validation. Because the naive baseline and the proposed pipeline use the same Qwen2.5-VL backbone, comparing them isolates the contribution of structural design rather than differences in underlying model capability.

Comparing these four configurations allows the evaluation to distinguish between general recognition ability and the added value of explicit document structure modeling for multi-page Writing-to-Learn artifacts. This comparison is particularly important because instructional usability depends not only on transcription quality, but also on correct attribution, section integrity, and student-level reconstruction.

6.5 Alignment with Writing-to-Learn Pedagogy

The evaluation design is intentionally aligned with the goals of Writing-to-Learn pedagogy. Because learning plans, evaluations, and reflections are designed to capture students' reasoning processes, planning strategies, and self-assessments, the evaluation focuses on whether those processes remain interpretable after transcription. Minor spelling errors are tolerated, whereas errors that distort meaning, attribution, or document structure are not. This framing ensures that evaluation outcomes are directly relevant to instructional decision-making rather than abstract technical benchmarks. In this sense, the evaluation criteria are designed not only to assess transcription quality, but also to determine whether the resulting text remains usable for pedagogically meaningful interpretation of student thinking and reflection.

6.6 Student Self-Evaluation and Instructor Qualitative Validation

To complement the protocol-level alignment of the technical evaluation with Writing-to-Learn pedagogy, a student self-evaluation was conducted to examine whether the AI-generated transcripts preserved the intended meaning of handwritten Writing-to-Learn (WTL) responses and whether handwriting was perceived as supportive of thinking and reflection. Handwritten WTL artifacts from two undergraduate engineering courses were used in this follow-up evaluation, yielding a total of 24 student submissions. In the subsequent semester, students were invited by email to voluntarily review their original handwritten responses together with the corresponding AI-generated transcripts and complete a short survey. Participation was voluntary, refusal involved no penalty or loss of benefits, and responses were used only for evaluation and research purposes. A total of seven students responded.

Table 1. Student self-evaluation survey instrument

Item	Construct	Survey Question	Response Format
Q1	Transcription Fidelity	To what extent do the transcribed texts represent the intent and meaning of your original handwritten responses?	5-point scale: 1 = Does not represent intent; 5 = Almost perfectly represents intent
Q2	Thinking Depth	Compared with typing a response, to what extent did writing by hand help you think more carefully?	5-point scale: 1 = Much less than typing; 5 = Much more than typing
Q3	Reflection & Awareness	Writing my response by hand helped me reflect more on what I understood (or did not understand).	5-point Likert scale: 1 = Strongly disagree; 5 = Strongly agree
Q4	Cognitive Engagement	Compared with typing, handwriting required me to slow down and think more about what I wanted to express.	5-point Likert scale: 1 = Strongly disagree; 5 = Strongly agree
Q5	Error Identification	If the transcript does not fully represent your handwritten response, please indicate the main issue or issues you observe.	Multiple selection / short descriptive response: Missing words or phrases; Incorrect words; Sentence structure or grammar issues; Meaning distortion; Handwriting was unclear; No issue / transcript acceptable

The student self-evaluation survey was administered through Google Forms and included five items. As summarized in Table 1, the first item asked students to rate how well the transcribed text represented the intent and meaning of their original handwritten responses. The next three items examined the perceived pedagogical effects of handwriting by asking whether handwriting supported more careful thinking, greater reflection on understanding, and slower, more deliberate expression compared with typing. The final item asked students to identify any main transcription issues, such as missing words, incorrect words, sentence structure or grammar problems, meaning distortion, or unclear handwriting. This five-item structure was intentionally designed to keep student burden low while providing both technical validation of transcript fidelity and initial evidence of the perceived educational value of handwriting in the WTL context.

In addition to the student self-evaluation, an instructor qualitative validation was conducted. The course instructor reviewed a subset of original handwritten responses together with the corresponding AI-generated transcripts and the research assistant–generated reference transcripts used in the technical evaluation. This review was intended to provide an instructional perspective on whether the generated transcripts adequately represented the handwritten responses and remained suitable for interpretation in the WTL context. Because this review was not collected as a separate scored dataset, it is treated as qualitative validation rather than as a formal quantitative measure. This added layer extends the technical evaluation by incorporating both end-user and instructional perspectives on semantic fidelity, pedagogical usability, and the value of handwriting for reflective learning.

7. Results and Validation

This section reports the quantitative transcription results, qualitative error patterns, and human-centered validation evidence used to evaluate the proposed handwriting-to-text pipeline.

7.1 Quantitative Comparison Across Models

Table 2 summarizes structured transcription accuracy across several representative approaches, including traditional OCR, OCR combined with vision–language model post-processing, an unconstrained vision–language baseline, and the proposed structured pipeline. Accuracy is measured at the student-submission level, meaning that a submission is counted as correct only if all required conditions are satisfied. These conditions include correct metadata association, correct grouping of multiple pages, preservation of semantic meaning, and absence of header leakage into student responses.

Table 2. Structured transcription accuracy comparison across different sample sizes

Method	5 Samples	10 Samples	20 Samples
Tesseract OCR	20%	35%	34%
Tesseract + Qwen2.5-VL Post-Processing	35%	46%	45%
Qwen2.5-VL (Naive Page-wise Prompting)	82.0%	83.6%	86.1%
Ours: Structured VLM Pipeline	**95.0%**	**97.1%**	**96.4%**

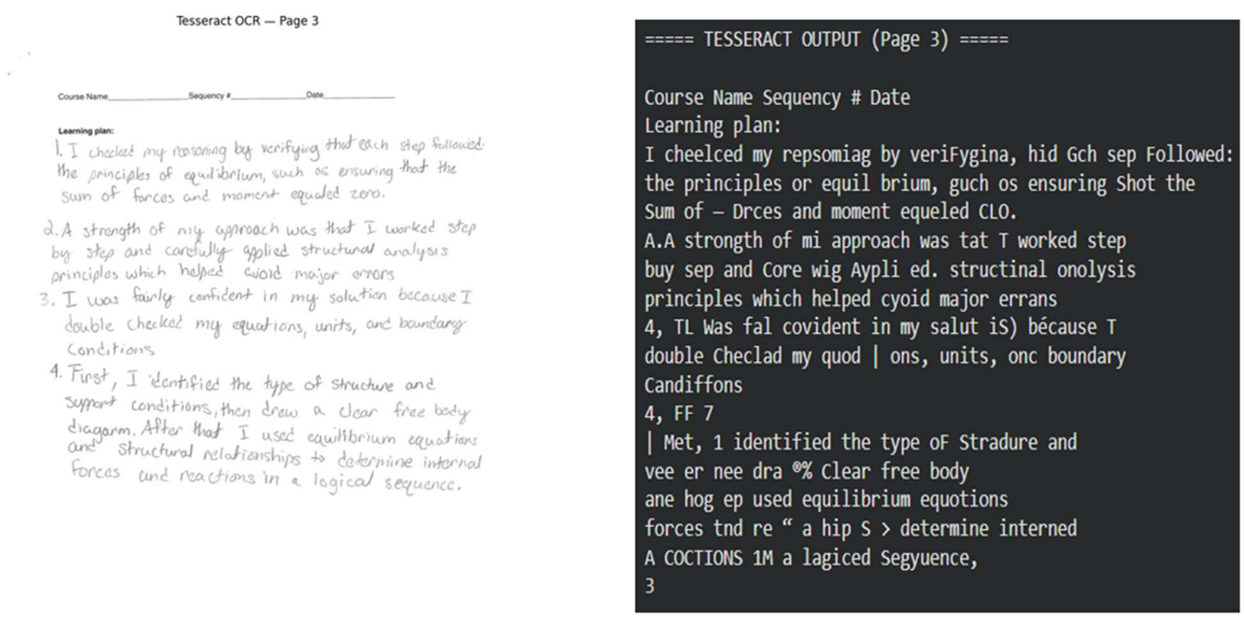
The results show a clear pattern. Traditional OCR performs poorly on handwritten student work, particularly for multi-page submissions. Errors such as missing text, incorrect character recognition, and failure to associate pages belonging to the same student frequently render the output unusable for instructional purposes. Adding Qwen2.5-VL post-processing improves formatting consistency but does not compensate for fundamental OCR failures, because missing or misrecognized content cannot be reliably reconstructed.

Vision–language models substantially improve transcription quality by directly interpreting handwritten images. However, when applied without explicit structural constraints, the naive Qwen2.5-VL baseline still introduces errors that are critical in educational contexts, such as mixing metadata with student responses or disordering pages within a submission. By contrast, the proposed structured pipeline achieves the highest accuracy across all evaluated sample sizes in Table 2.

Importantly, the naive Qwen2.5-VL baseline and the proposed structured pipeline use the same underlying model backbone. Thus, the observed performance gain is attributable not to a stronger model, but to the introduction of structural design features, including identity-aware prompting, metadata separation, and student-level multi-page grouping. These results indicate that reliable transcription of multi-page handwritten Writing-to-Learn artifacts depends not only on model capability, but critically on document-level structural design.

7.2 Qualitative Error Patterns and Visual Evidence

Figure 2 provides qualitative examples that help explain the quantitative trends observed in Table 2. The figure illustrates representative transcription outputs generated by different approaches when applied to the same handwritten student submissions.



(a) Original handwritten student response (b) Tesseract OCR transcription output

Figure 2: Qualitative example illustrating failure modes of traditional OCR on handwritten student responses. Tesseract frequently produces incomplete transcriptions, misrecognized

characters, and mixed header/body content, leading to incorrect metadata extraction and unusable structured outputs.

The dominant failure modes differ across method families. Traditional OCR outputs shown in Figure 2 contain fragmented or corrupted text, misrecognized characters, and merged header and body content that obscure student reasoning and often lead to incorrect metadata extraction. Tesseract with Qwen2.5-VL post-processing recovers some formatting and fluency, but it cannot reliably reconstruct content that was lost or severely distorted during the initial OCR stage. The naive Qwen2.5-VL baseline achieves substantially better transcription quality, but when structural constraints are not enforced it still produces errors that are critical for instructional use, including metadata leakage into response bodies and conflation of pages across students. By contrast, the proposed structured pipeline preserves both document structure and semantic intent by design, maintaining the separation between metadata and student response content while correctly reconstructing multi-page submissions as coherent student-level records.

These qualitative examples reinforce the quantitative findings in Table 2 by showing that the structured pipeline's advantage arises not only from stronger transcription performance, but from the prevention of failure modes that make handwritten Writing-to-Learn responses instructionally unusable. In particular, the visual evidence shows why document-level structural modeling is necessary for reliable transcription of multi-page reflective writing.

7.3 Student Self-Evaluation and Instructor Qualitative Validation

As described in Section 6.6, the student self-evaluation was conducted to complement the technical findings by examining whether the generated transcripts preserved the intended meaning of handwritten Writing-to-Learn responses and whether handwriting was perceived as supportive of thinking and reflection. The results are summarized in Table 3. Overall, the student responses indicate generally strong perceived transcription fidelity. The mean rating for transcript fidelity was 4.14 out of 5.00, and 85.7% of respondents rated the transcripts 4 or higher, suggesting that most students perceived the generated text as representing the intended meaning of their handwritten responses reasonably well.

Table 3. Student self-evaluation results

Item	Measure	Result	Rate 4 or Higher
Q1	Transcription Fidelity	Mean = 4.14 / 5.00	85.7%
Q2	Thinking Depth	Mean = 3.86 / 5.00	85.7%
Q3	Reflection & Awareness	Mean = 4.00 / 5.00	71.4%
Q4	Cognitive Engagement	Mean = 4.14 / 5.00	85.7%
Q5	If not accurate, due to what reason?	No issue / acceptable transcript (n = 4); Missing words or phrases (n = 1); Incorrect word(s) (n = 1); Sentence structure or grammar issues (n = 1); Meaning distortion (n = 0); Handwriting was unclear (n = 1)	

The ratings also provide preliminary evidence regarding the perceived pedagogical value of handwriting. As shown in Table 3, the mean rating for thinking depth was 3.86, with 85.7% of respondents rating it 4 or higher, suggesting that most students perceived handwriting as supporting more careful thinking compared with typing. The mean rating for reflection and awareness was 4.00, with 71.4% rating it 4 or higher, indicating generally positive agreement that handwriting helped students reflect on what they understood or did not understand. The mean rating for cognitive engagement was 4.14, with 85.7% rating it 4 or higher, suggesting that handwriting encouraged slower and more deliberate expression.

Responses to the issue-identification item were analyzed descriptively. Four of the seven respondents reported no meaningful transcription problem. Among the remaining responses, the reported issues were limited and minor, including missing words or phrases ($n = 1$), incorrect word choice ($n = 1$), sentence structure or grammar issues ($n = 1$), and unclear handwriting ($n = 1$). No respondent identified meaning distortion as a reported issue. Because one respondent could report more than one issue, these counts are not mutually exclusive. Taken together, the issue reports suggest that transcription errors were generally limited and did not substantially alter the interpretability of the responses.

In addition to the student self-evaluation, the instructor qualitative validation was consistent with these findings. After reviewing a subset of original handwritten responses, AI-generated transcripts, and research assistant reference transcripts, the instructor judged the generated transcripts to be generally consistent with the handwritten responses and suitable for instructional interpretation in the Writing-to-Learn context. Although this review was qualitative rather than score-based, it provides an additional instructional perspective on semantic fidelity and pedagogical usability. Taken together, the student ratings summarized in Table 3 and the instructor validation complement the technical findings reported in Sections 7.1 and 7.2 and provide preliminary human-centered support for the instructional usability of the proposed pipeline.

8. Discussion

The results of this study demonstrate the feasibility of using AI-based handwriting-to-text transcription to support handwritten Writing-to-Learn (WTL) artifacts in undergraduate engineering courses. Across learning plans, learning evaluations, and learning reflections, the proposed structured pipeline converted multi-page handwritten submissions into organized, machine-readable text while preserving document structure and semantic content at a level suitable for instructional use. These findings indicate that handwritten reflective responses can be incorporated into digital instructional workflows without requiring instructors to manually process the original handwritten pages.

From an educational perspective, the main contribution of this work is not simply the development of a transcription system, but the demonstration that reflective handwritten WTL artifacts can be supported at scale in the generative-AI era. This contribution is important because the artifacts examined here are designed to capture planning, self-monitoring, and reflection rather than correctness alone. By making such artifacts more manageable for instructors, the proposed pipeline helps preserve the pedagogical value of WTL while reducing one of its major practical barriers.

The evaluation approach further underscores the importance of aligning technical assessment with instructional use. Rather than emphasizing character-level recognition accuracy alone, the study evaluates transcription at the student-submission level, incorporating correct page grouping, metadata association, and preservation of semantic meaning. The results show that the structured design of the pipeline matters substantially: because the naive Qwen2.5-VL baseline and the proposed structured pipeline use the same underlying backbone, the strongest gains are attributable not to a stronger model, but to the combination of identity-aware prompting, metadata separation, and student-level grouping. For handwritten WTL artifacts, reliable instructional use depends on preserving meaning and structure together.

The findings also extend beyond technical performance by providing preliminary human-centered support for instructional usability. Student self-evaluation suggests that the generated transcripts generally preserved intended meaning, while the added student perception items indicate that handwriting was still viewed as supporting careful thinking and reflection. In parallel, the instructor qualitative validation suggests that the generated transcripts were broadly consistent with the handwritten responses and suitable for instructional interpretation. Taken together, these results strengthen the argument that the system functions as enabling infrastructure for reflective learning rather than merely as a document-processing tool.

Finally, the study carries an important representational implication because the evaluation was conducted in an HBCU instructional context. Although a formal subgroup fairness analysis is beyond the scope of the current study, the system was evaluated on authentic student writing from a population that is underrepresented in existing handwriting-recognition benchmarks. This context adds broader significance to the findings by providing initial evidence that AI-based handwriting transcription can be examined in more diverse and educationally meaningful settings, rather than only on standard benchmark datasets.

9. Limitations and Future Work

Several limitations define the current scope of this study. First, the analysis is limited to handwritten textual responses. Engineering artifacts that depend heavily on diagrams, equations, or symbolic representations were not examined, and the findings should not be generalized to those forms of work. This boundary is consistent with the current pipeline design, which focuses on handwritten text and structured multi-page WTL artifacts rather than multimodal engineering documents.

Second, the transcription pipeline was evaluated as an enabling layer for instructional use, not as an automated grading or feedback system. The technical evaluation established strong structured transcription performance, and the added student self-evaluation and instructor qualitative validation provided initial evidence of semantic fidelity and instructional usability. However, these human-centered findings remain preliminary. The student follow-up involved a small voluntary sample, and the instructor review was qualitative rather than a separate scored dataset. In addition, because semantic fidelity was judged through human verification of handwritten responses and generated transcripts, the analysis may still be affected by interpretive bias in how meaning preservation was determined, even though consistent transcription rules and instructor review were used to strengthen validity.

Third, the data were drawn from a limited number of courses within a single institutional context. Although this setting is valuable because it reflects authentic classroom conditions and contributes evidence from an HBCU instructional environment, broader deployment across institutions, disciplines, and student populations is still needed to assess robustness, equity, and generalizability at scale. The present study therefore provides an initial rather than definitive account of how well such a pipeline may function across diverse educational settings.

Although the proposed structured pipeline substantially reduced document-level errors, several edge cases remained challenging. These included responses with heavily corrected or overwritten words, faint or low-contrast handwriting, irregular spacing between lines, and pages where header information was written too close to response content. These cases suggest that future improvements should focus not only on model recognition accuracy, but also on preprocessing, layout normalization, and confidence-based human review.

Future work should proceed in several directions. One priority is to expand the student-centered evaluation with larger samples and more systematic instructor-based validation. A second is to examine broader inter-rater and cross-course consistency in judgments of transcript adequacy for WTL use. A third is to extend the pipeline to multimodal artifacts that combine handwritten text with diagrams, equations, or symbolic representations, including edge cases such as heavily corrected pages or highly stylized cursive. Finally, future work should continue integrating the handwriting-to-text pipeline with rubric-based, AI-powered assessment tools so that handwritten WTL responses can support not only transcription and instructional review but also more scalable automatic assessment and adaptive learning support.

10. Conclusion

This paper presented a study of AI-based handwriting-to-text transcription to support handwritten Writing-to-Learn (WTL) activities in undergraduate engineering courses. Using authentic classroom data, the results show that the proposed structured pipeline can reliably convert multi-page handwritten WTL artifacts into organized, machine-readable text suitable for instructional use. Because the naive Qwen2.5-VL baseline and the structured pipeline use the same underlying model backbone, the observed performance gains indicate that reliable transcription of handwritten WTL artifacts depends not only on model capability, but critically on structural design choices such as identity-aware prompting, metadata separation, and student-level grouping.

From an educational perspective, the main contribution of this study is to show that reflective handwritten WTL artifacts—not only short-answer or correctness-oriented responses—can be supported at scale through AI-assisted transcription without losing their pedagogical relevance. The technical findings, together with the added student self-evaluation and instructor qualitative validation, provide initial evidence that the generated transcripts generally preserve intended meaning and remain usable for reflective instructional interpretation. The findings also suggest that handwriting continues to support careful thinking and reflection when paired with AI-assisted transcription.

The proposed approach is intended as enabling infrastructure rather than as an automated grading system by itself. Its value lies in allowing instructors to continue using handwritten responses as authentic WTL artifacts while reducing the practical burden of transcription and organization. In addition, this handwriting-to-text pipeline establishes a foundation for future integration with rubric-based, AI-powered automatic assessment systems currently under development. Once integrated, the transcribed handwritten responses could support more efficient assessment of student writing, assist instructors in evaluating student thinking, and provide a basis for adaptive feedback and learning support. Taken together, these findings provide initial support that AI-assisted transcription can help sustain the pedagogical value of handwritten WTL in the generative-AI era while also enabling future AI-supported assessment and adaptive instructional use.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant Nos. NSF/CNS-2302686 and NSF/EES-2306566. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Boston, MA: Center for Curriculum Redesign, 2019. Available: https://curriculumredesign.org

[2] R. Luckin, W. Holmes, M. Griffiths, and L. B. Forcier, *Intelligence Unleashed: An Argument for AI in Education*. London, UK: Pearson Education, 2016.

[3] P. A. Mueller and D. M. Oppenheimer, "The pen is mightier than the keyboard: Advantages of longhand over laptop note taking," *Psychological Science*, vol. 25, no. 6, pp. 1159–1168, 2014.

[4] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. on Machine Learning (ICML)*, New York, NY, USA: ACM, 2006, pp. 369–376.

[5] C. Wigington, S. Stewart, B. Davis, B. Barrett, B. Price, and S. Cohen, "Data augmentation for recognition of handwritten words and lines using a CNN-LSTM network," in *Proc. 15th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR)*, 2018, pp. 639–645.

[6] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-VL technical report," arXiv preprint arXiv:2502.13923, 2025.

[7] R. Smith, "An overview of the Tesseract OCR engine," in *Proc. 9th Int. Conf. on Document Analysis and Recognition (ICDAR)*, 2007, pp. 629–633.

[8] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2020. Available: https://arxiv.org/abs/1904.09675

[9] R. S. Baker and A. Hawn, "Algorithmic bias in education," *International Journal of Artificial Intelligence in Education*, vol. 31, no. 2, pp. 105–122, 2021.

[10] Y. Li *et al*., "Vision-language models for document understanding: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. Available: https://ieeexplore.ieee.org/document/10105924

[11] Amazon Web Services, "Amazon EC2 and scalable machine learning inference," 2024. Available: https://aws.amazon.com/ec2/

[12] J. Wei *et al*., "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b0a3b6c1-Paper.pdf