

**Paper Type: Original Article**

# Catalyst: A Human-Supervised AI Framework for End-to-End Handwritten Assessment and Personalized Learning Support

Mohamed Ahmed <sup>1</sup> , Abdullah Samir <sup>1</sup> , Mostafa Zaki <sup>1</sup> , Mahmoud Khaled <sup>1</sup> , Mohamed Wael <sup>1</sup>   
and Abdelrhman Fathy <sup>1,\*</sup> 

<sup>1</sup> Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Sharqiyah, Egypt;  
Emails: [elmalkymohamed278@gmail.com](mailto:elmalkymohamed278@gmail.com), [abdullhsamir2@gmail.com](mailto:abdullhsamir2@gmail.com), [mostafazaki026@gmail.com](mailto:mostafazaki026@gmail.com),  
[mahmoudkhalid0987@gmail.com](mailto:mahmoudkhalid0987@gmail.com), [mwael4128@gmail.com](mailto:mwael4128@gmail.com), [asawmaa@fci.zu.edu.eg](mailto:asawmaa@fci.zu.edu.eg).

**Received:** 12 May 2026**Revised:** 23 Jul 2026**Accepted:** 04 Oct 2026**Published:** 07 Oct 2026**Abstract**

Artificial-intelligence applications in education are increasingly expected to support assessment, feedback, and learning analytics, yet many prototypes evaluate a single model without showing how it can be embedded in a coherent educational workflow. This paper presents Catalyst, an integrated educational platform that connects handwritten-answer processing, assessment management, teacher verification, learner feedback, and performance-oriented support within a modular client-server architecture. The system uses a Flutter mobile client, a Spring Boot backend, PostgreSQL persistence, a two-stage document-processing pipeline that combines Differentiable Binarization-based text detection with TrOCR recognition, and a T5-based text-to-text service for content summarization and explanation. The principal technical evaluation focuses on handwriting recognition because it provides an objective, ground-truth-based measure. Word Error Rate (WER) is used throughout without converting WER into an artificial “accuracy” score. The reported evaluation yields WER = 0.0090 for single-line handwriting and WER = 0.1925 for multi-line document processing, showing that layout analysis and line extraction are the dominant sources of degradation at page level. In the recorded model comparison, TrOCR-Large and TrOCR-Base produced substantially lower WER than Tesseract, EasyOCR, and Donut under the project evaluation configuration. Beyond model performance, Catalyst separates machine transcription from final academic authorization: objective items may be scored automatically, whereas written responses remain teacher-verifiable before publication. The resulting contribution is an end-to-end, human-supervised architecture that links document AI to actionable educational workflows rather than treating recognition, grading, analytics, and learning support as disconnected functions.

**Keywords:** Artificial Intelligence In Education; Handwritten Text Recognition; TrOCR; OCR; Learning Analytics; Automated Assessment; Human-in-the-Loop; Educational Technology.

## 1 | Introduction

Artificial intelligence in education (AIED) now spans adaptive learning, intelligent assessment, learner modelling, administrative support, and data-informed feedback. Recent reviews describe the field as technically mature enough to support multiple educational functions, but also emphasize that educational value depends on how models are embedded in pedagogy, workflow, and human decision making rather than



Corresponding Author: [asawmaa@fci.zu.edu.eg](mailto:asawmaa@fci.zu.edu.eg)



<https://doi.org/10.61356/j.scin.2026.3677>



Licensee **SciNexuses**. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

on model novelty alone [1]. Earlier reviews similarly showed that AIED research has often concentrated on technical capability while giving less attention to educators, classroom integration, and operational constraints [2]. For learning platforms, this creates a practical design problem: an accurate model is not sufficient unless its outputs can be securely incorporated into assessment, verification, feedback, and follow-up processes.

Conventional e-learning systems are effective at content distribution, communication, enrollment, and grade management, yet empirical work on e-learning success shows that system quality, information quality, service quality, and user experience jointly influence educational usefulness [3]. At the same time, automatic assessment of open-ended responses remains an active research area because large-scale grading is costly and technically difficult; recent systematic evidence documents rapid progress in NLP-based assessment but also stresses the need for transparent evaluation and appropriate human oversight [4]. These findings motivate an architecture that treats artificial intelligence as one layer of an educational system rather than as a stand-alone replacement for instructors.

Catalyst addresses this integration problem by linking five functions in one workflow: (i) handwritten document ingestion and recognition, (ii) assessment creation and timed delivery, (iii) teacher-controlled verification of written responses and scores, (iv) learner-facing explanations and summaries, and (v) performance-oriented recommendations derived from assessment history. The design is deliberately human-supervised. AI output can accelerate transcription, content preparation, and feedback generation, but high-stakes academic decisions remain reviewable by the teacher, consistent with human-centred AIED principles that emphasize agency, reliability, and accountability [5].

## 1.1 | Research Gap and Scientific Contribution

The scientific gap addressed in this work is not the absence of OCR, language models, learning analytics, or learning-management systems as separate technologies. Each of these areas is already well established. The gap is the lack of a clearly specified engineering pathway that connects handwritten educational evidence to a controlled assessment lifecycle and then to learner support while preserving human authorization. Recent learning-analytics research has likewise argued that useful dashboards should move beyond passive display toward interpretable feedback and action [6], and newer work has explored interactive, pedagogically informed dashboards rather than static visualization alone [7]. Catalyst operationalizes that direction at the assessment-workflow level.

**Table 1.** Research gap and scientific contribution of Catalyst.

| Research stream                     | Typical capability                                                                                     | Unresolved integration issue                                                                     | Catalyst contribution                                                                                                |
|-------------------------------------|--------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|
| <b>Handwritten text recognition</b> | Transcribes cropped text images or document regions [8–10].                                            | Recognition output is often evaluated apart from the assessment workflow.                        | Two-stage page-to-line recognition is connected directly to exam ingestion and teacher verification.                 |
| <b>Automatic assessment</b>         | Scores or analyzes text responses; literature spans rule-based, ML, and language-model approaches [4]. | Automatic scoring can obscure the distinction between machine output and academic authorization. | Written responses remain reviewable; objective scoring and transcription are separated from final score publication. |
| <b>Learning analytics</b>           | Visualizes or analyzes learner activity and performance [5,6].                                         | Analytics may remain descriptive rather than action-oriented.                                    | Weakness summaries are linked to revision topics, practice, and teacher follow-up workflows.                         |
| <b>AI-assisted learning support</b> | Generates explanations, summaries, or tutoring interactions [1,7,11].                                  | Generative support may be disconnected from verified assessment evidence.                        | The content-support service is placed downstream of authenticated student,                                           |

|                                       |                                                                          |                                                                                 |                                                                                                          |
|---------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
|                                       |                                                                          |                                                                                 | lesson, and assessment context.                                                                          |
| <b>Educational system engineering</b> | Provides LMS functions such as class, content, and grade management [3]. | AI components are frequently external or loosely coupled to transaction states. | A modular backend coordinates AI calls, exam states, role-based access, persistence, and recovery logic. |

Table 1 positions the contribution at the intersection of document intelligence, assessment control, and actionable learning support. The paper therefore evaluates the OCR component quantitatively while documenting the surrounding architecture and safeguards as a reproducible system design.

## 1.2 | Research Questions

The study is organized around three questions. RQ1 asks whether a line-oriented TrOCR pipeline can provide low word error on handwritten educational text. RQ2 asks how recognition changes when the same recognition concept is extended from isolated lines to multi-line document images that require detection, ordering, and cropping. RQ3 asks how the selected recognition model compares with alternative OCR approaches under the project evaluation configuration. A fourth, system-level question concerns how recognition output can be integrated into an assessment workflow without treating machine transcription as an irreversible grading decision.

## 2 | Related Work

### 2.1 | Artificial Intelligence, Assessment, and Learning Analytics

AIED research has broadened from tutoring and prediction to intelligent assessment, management, personalized support, and emerging generative products [1]. Within assessment, a systematic review of automatic evaluation of text-based responses identified a wide range of input, processing, and output designs, confirming both the scalability opportunity and the need to align technical evaluation with educational purpose [4]. Assessment-for-learning research similarly emphasizes that feedback should contribute to subsequent learning rather than merely produce a score [16]. These observations inform Catalyst’s separation between assessment processing, teacher verification, and learner-facing support.

Learning analytics introduces another relevant layer. Human-centred reviews argue that useful educational AI should maintain an explicit balance between automation and stakeholder control, particularly where trust, safety, and reliability affect adoption [5]. Systematic evidence on learning analytics dashboards also shows a transition from purely descriptive interfaces toward designs that support reflection and action [6]. Catalyst follows this direction by using performance data to organize targeted follow-up while keeping the teacher able to inspect the underlying assessment evidence.

### 2.2 | Handwritten Text Recognition and Document Processing

Handwritten text recognition (HTR) is a sequence-transcription problem in which visual text is converted to character or token sequences. TrOCR introduced an end-to-end Transformer-based recognizer that combines a vision Transformer encoder with a text Transformer decoder and demonstrated strong performance on handwritten and printed recognition benchmarks [8]. The IAM Handwriting Database is a standard English handwriting resource containing scanned forms, lines, words, and sentence-level annotations from many writers, making it suitable for line-level HTR evaluation [9].

Full-page educational documents require an additional localization stage before recognition. Differentiable Binarization (DB) is a segmentation-based text-detection method that learns an adaptive binarization process and can produce text regions efficiently [10]. Catalyst uses this principle to detect text regions, sort them in

reading order, crop line candidates, and pass each crop to TrOCR. This decomposition is important because the recognizer is optimized for text-region input rather than arbitrary full-page layouts.

The comparison models represent different recognition philosophies. Tesseract is a mature OCR engine with a long history in document recognition [14], whereas Donut was proposed as an OCR-free Transformer architecture for visual document understanding [13]. Their inclusion in the project comparison provides useful contrast, but the present study interprets those results only within the recorded evaluation configuration; it does not claim that one model is universally superior across document domains.

## 2.3 | Text-to-Text Educational Support

T5 formulates diverse NLP tasks within a unified text-to-text framework [11], building on Transformer attention mechanisms [12]. In Catalyst, a T5-Large service is used as a content-support component for summarization and simplified explanations. The present experimental section does not claim a separate benchmark of educational text generation quality; the module is documented as part of the integrated system architecture, while quantitative evaluation is reserved for the handwriting pipeline where ground-truth transcriptions permit objective error measurement.

## 3 | Catalyst System Design

### 3.1 | Layered Architecture

Catalyst follows a modular client-server architecture. The Flutter mobile client communicates with a Spring Boot backend through REST endpoints. The backend separates request handling, domain logic, persistence, and AI-service orchestration so that assessment state changes are controlled by application logic rather than by the mobile client. Figure 1 shows the principal presentation, service, data-access, and external-service relationships.

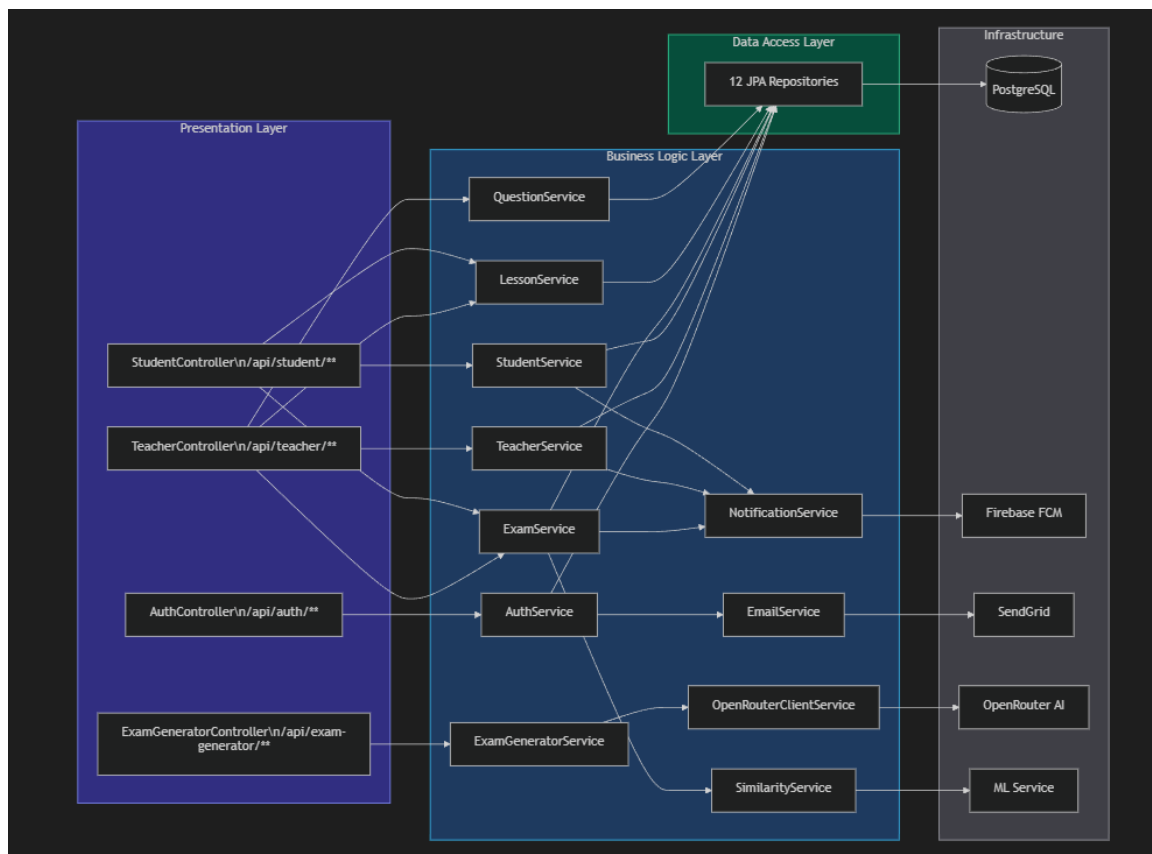


Figure 1. Layered Catalyst software architecture and service integration.

The controller layer exposes role-aware endpoints for authentication, classes, lessons, examinations, results, and AI requests. The service layer implements the application rules, including exam-state transitions, score publication, session recovery, and calls to the AI services. Repository components isolate database operations through Spring Data JPA, while PostgreSQL stores persistent entities and relationships. This separation reduces coupling between user-interface behavior and core academic rules.

### 3.2 | Data Model and Academic State

The database schema, shown in Figure 2, organizes users, classes, lessons, examinations, questions, student attempts, results, and supporting records into relational entities. The data model is designed so that an exam attempt, its recognized or manually entered responses, teacher review state, and final published result can be tracked as distinct records rather than collapsed into a single opaque score.

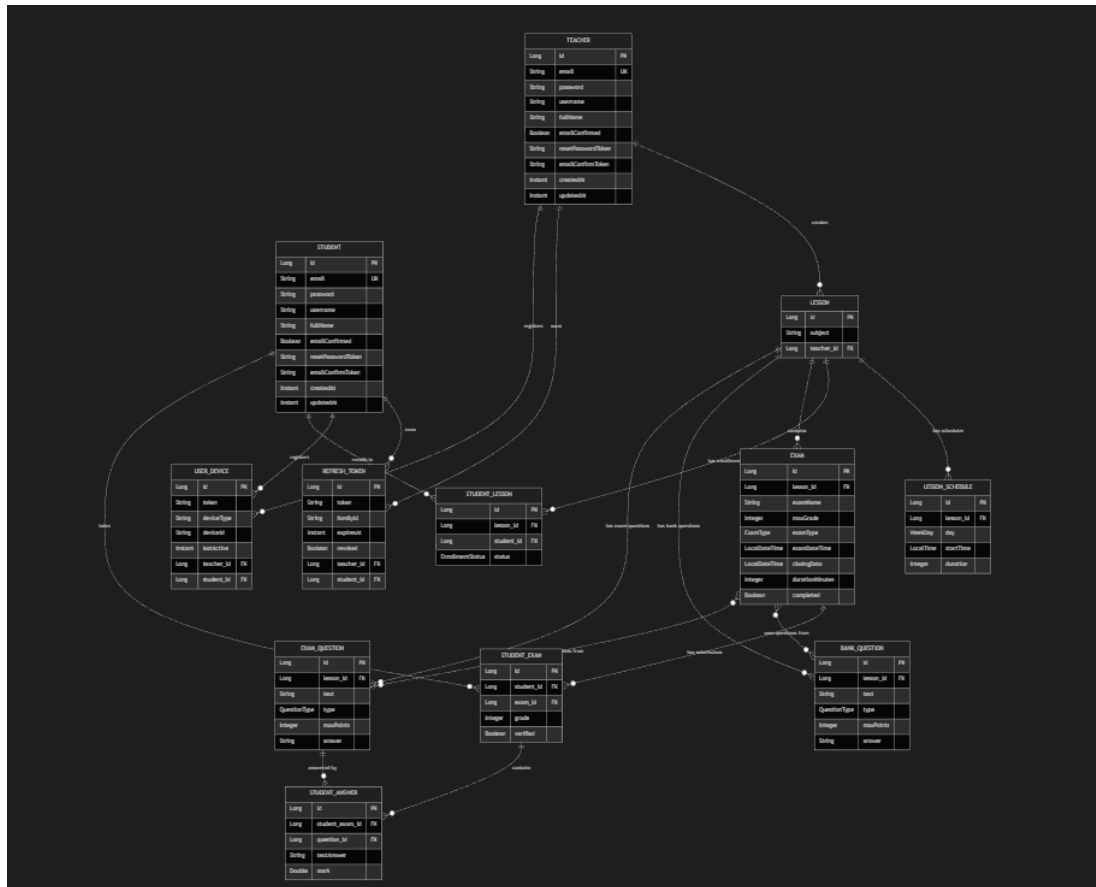


Figure 2. Relational database schema used by Catalyst.

### 3.3 | Technology Stack and Functional Modules

The implementation combines a Java/Spring Boot backend, PostgreSQL persistence, a Flutter/Dart mobile client, and Python-based AI services. The handwriting service uses image preprocessing, DB-style line detection through an OpenCV DNN model, and the pretrained microsoft/trocr-large-handwritten recognizer. The content-support service uses T5-Large in a text-to-text configuration. Table 2 summarizes the functional role of each major module and clarifies which outputs are automatically produced and which remain subject to teacher review.

**Table 2.** Catalyst modules, outputs, and control points.

| Module                                 | Primary technology                     | Output                                                  | Human/control boundary                                                               |
|----------------------------------------|----------------------------------------|---------------------------------------------------------|--------------------------------------------------------------------------------------|
| <b>Authentication and role control</b> | Spring Security, token-based sessions  | Authorized user session and role-specific access        | Server validates identity and role before protected operations.                      |
| <b>Class and lesson management</b>     | Spring Boot + PostgreSQL               | Classes, enrollment, lessons, attendance records        | Teacher controls publication and class membership.                                   |
| <b>Handwriting recognition</b>         | DB text detection + TrOCR-Large [8,10] | Machine-readable transcription of handwritten text      | Transcription supports review; it is not treated as an irreversible final grade.     |
| <b>Assessment workflow</b>             | Backend exam services                  | Exam delivery, timing, attempt state, objective scoring | Server controls official timing and state; written-answer results remain reviewable. |
| <b>Content support</b>                 | T5-Large [11]                          | Summaries and simplified explanations                   | Generated content is support material, not authoritative academic judgment.          |
| <b>Performance support</b>             | Assessment-history analysis            | Weak-area summaries and recommended follow-up           | Teachers and students can inspect results and act on recommendations.                |

### 3.4 | Reliability and Assessment-Integrity Safeguards

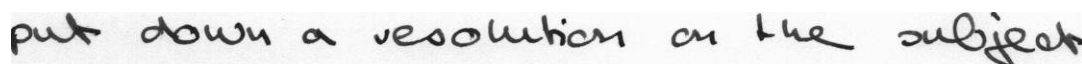
Reliability in an educational assessment platform depends on controlled state transitions rather than on model output alone. Catalyst therefore manages exam timing on the server, maintains explicit states such as upcoming, active, pending review, verified, and missed, and supports restoration of an interrupted attempt using the remaining valid server-side time. This prevents a local device clock from determining the official deadline and reduces the risk that an application restart creates a new attempt.

Role-based authorization separates teacher functions from student functions. Objective items can be evaluated automatically according to stored answer keys, whereas written responses can remain pending until teacher review. This human-in-the-loop boundary is consistent with current guidance that educational AI should preserve meaningful human responsibility, data protection, and context-appropriate oversight [5, 15].

## 4 | Methodology

### 4.1 | Data Sources

The recognition experiments use English handwritten text drawn from the IAM handwriting resource and a multi-line document set prepared for page-level processing. The IAM database is widely used for offline handwriting recognition and provides scanned forms together with segmented lines and transcriptions [9]. A representative line-level sample is shown in Figure 3. In the line-level pipeline, images are normalized before recognition; the project data loader uses line images resized to a fixed height of 128 pixels, and the TrOCR processor subsequently prepares model-compatible input.

**Figure 3.** Example of an IAM line-level handwriting sample [9].

The multi-line condition introduces document layout, labels, and multiple handwriting regions, as illustrated in Figure 4. Unlike isolated lines, these pages require detection of the relevant handwritten area, ordering of

text regions, and line-level cropping before recognition. The purpose of evaluating both conditions is therefore to distinguish recognizer quality from the additional error introduced by page segmentation and layout handling.

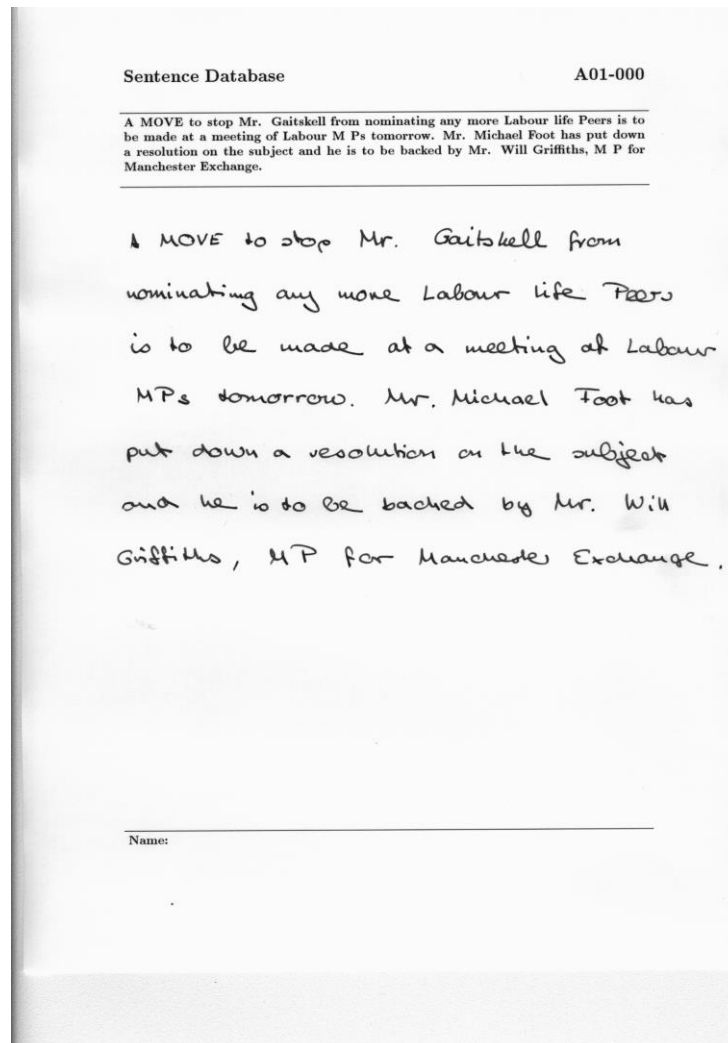


Figure 4. Representative multi-line handwritten document used for page-level processing.

## 4.2 | Preprocessing and Label Preparation

For line-level images, preprocessing converts the input to RGB, derives a grayscale representation for enhancement, applies automatic contrast adjustment and median filtering with a kernel size of 3, and prepares a  $384 \times 384$  model input. For the multi-line data, the pipeline additionally isolates the label area from the handwritten region. Grayscale conversion, Gaussian smoothing, Canny edge detection, morphological processing, and contour analysis are used to identify horizontal structures and separate the upper label region from handwriting. Figure 5 summarizes this preparation flow.

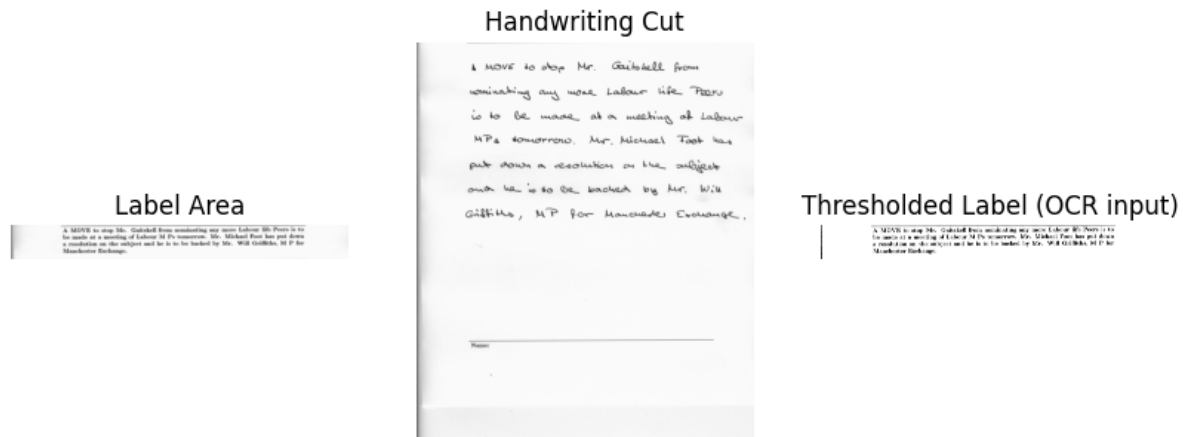


**Figure 5.** Multi-line preprocessing: label isolation, handwriting region extraction, and thresholded OCR label preparation.

The isolated label region is enlarged and binarized using Otsu thresholding before Tesseract OCR is applied to obtain an initial transcription. Samples with mismatches between visible handwriting and extracted labels are reviewed and corrected so that the evaluation compares recognition output with a consistent reference transcription. The resulting process addresses a common source of misleading OCR metrics: noisy or incomplete ground truth.

### 4.3 | Two-Stage Handwriting Recognition Pipeline

Catalyst processes full-page handwriting in two stages. First, a pretrained text detector based on Differentiable Binarization identifies text polygons. The implementation uses the `text_detection_en_ppocrv3` model through OpenCV DNN with a  $736 \times 736$  input, binary threshold 0.3, polygon threshold 0.5, unclip ratio 2.0, and a maximum of 200 candidate regions. Detected regions are sorted from top to bottom and cropped to isolate line-like text segments. Second, each crop is passed to `microsoft/trocr-large-handwritten`. TrOCR uses a vision encoder and autoregressive text decoder, following the architecture introduced by Li et al. [8]; the overall encoder-decoder principle is shown in Figure 6.

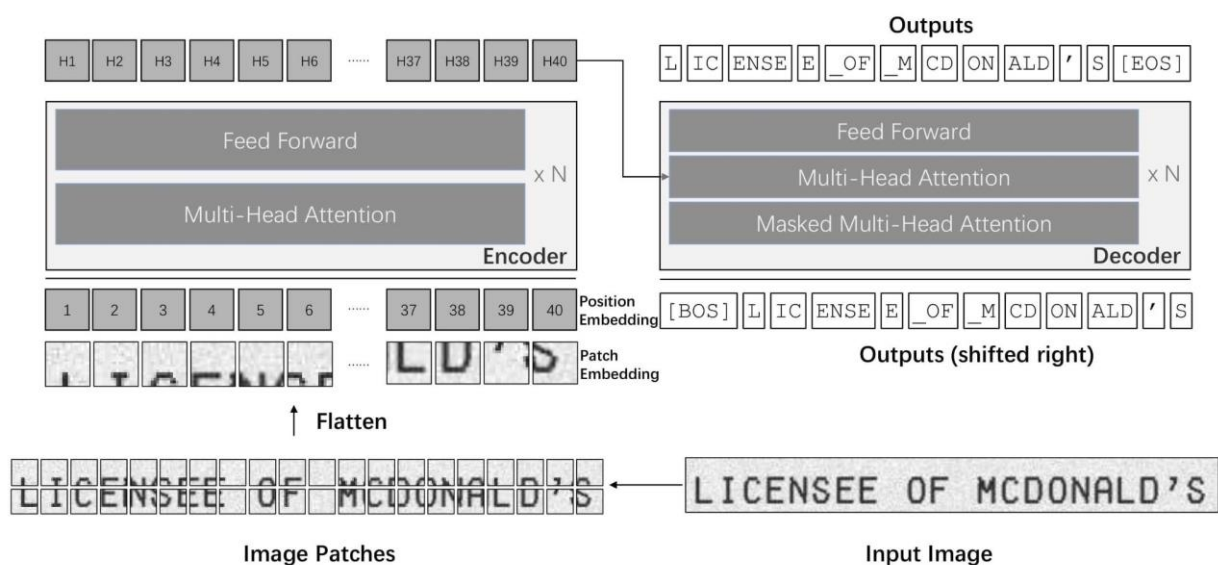


**Figure 6.** TrOCR vision-encoder/text-decoder recognition concept (adapted from Li et al. [8]).

The recognizer receives each RGB crop through the TrOCR processor, generates output token identifiers autoregressively, decodes them to text, and concatenates the line transcriptions in reading order. The present implementation uses the pretrained recognizer without additional TrOCR fine-tuning on the project dataset. This choice makes the experiment a test of transfer performance within the integrated pipeline rather than a report of a newly trained recognition model.

## 4.4 | Content-Support and Backend Integration

The content-support module uses T5-Large as a text-to-text service for lesson summarization and simplified explanations [11]. Requests are routed by the Spring Boot controller to the service layer, which invokes the Python AI service and stores the resulting output when required by the workflow. This module is architecturally independent of the handwriting recognizer so that document transcription failures do not directly alter language-model behavior and each component can be updated separately.

## 4.5 | Evaluation Protocol and Metric

The primary evaluation metric is Word Error Rate (WER), computed from the minimum word-level edit path between the reference transcription and the recognized sequence. If  $S$  denotes substitutions,  $D$  deletions,  $I$  insertions, and  $N$  the number of words in the reference, then WER is defined as follows:

$$WER = \frac{S + D + I}{N}$$

$WER = 0$  represents an exact word sequence, while larger values indicate more transcription errors. Importantly, WER is not bounded above by 1 because insertion errors can make  $S + D + I$  exceed  $N$ . For that reason, the revised evaluation reports WER directly and does not convert  $1 - WER$  into a quantity labelled “accuracy.” This resolves the interpretive problem that previously produced a negative pseudo-accuracy for EasyOCR. The study treats the reported WER values as engineering performance indicators for the implemented system configuration rather than as estimates of a universal population effect.

# 5 | Results and Discussion

## 5.1 | Line-Level and Multi-Line Recognition

Table 3 reports the principal recognition result and directly answers RQ1 and RQ2. Single-line handwriting produces  $WER = 0.0090$ , whereas the multi-line pipeline produces  $WER = 0.1925$ . The difference is substantial and is consistent with the additional detection, cropping, ordering, and layout operations required by page-level processing. Because the same general recognition approach is used downstream, the result identifies document decomposition as a major source of system-level error rather than attributing all degradation to the TrOCR decoder itself.

**Table 3.** Handwriting-recognition performance under the two input conditions.

| Evaluation condition         | WER ↓  | Interpretation                                                                   |
|------------------------------|--------|----------------------------------------------------------------------------------|
| Single-line handwriting      | 0.0090 | Very low word-level edit error for isolated line input.                          |
| Multi-line document pipeline | 0.1925 | Higher error after page detection, line extraction, and sequence reconstruction. |

The engineering implication of Table 3 is that improving the next version of Catalyst should prioritize page segmentation, line-order reconstruction, and robustness to irregular handwriting before increasing recognition-model size. A stronger recognizer cannot recover text that was missed, merged, or incorrectly cropped upstream.

## 5.2 | Comparison with Alternative OCR Models

RQ3 is addressed by the recorded model comparison shown in Figure 7 and Table 4. The revised comparison uses only WER because it is the metric consistently defined for all five outputs. Under this project configuration, TrOCR-Large records WER = 0.0090 and TrOCR-Base WER = 0.0267, whereas Tesseract, Donut, and EasyOCR produce much larger edit rates. The Donut result should be interpreted in light of its original purpose as an OCR-free visual-document-understanding model rather than a specialized handwritten line recognizer [13]. Likewise, Tesseract is a general OCR engine and is not designed specifically for unconstrained handwriting [14].

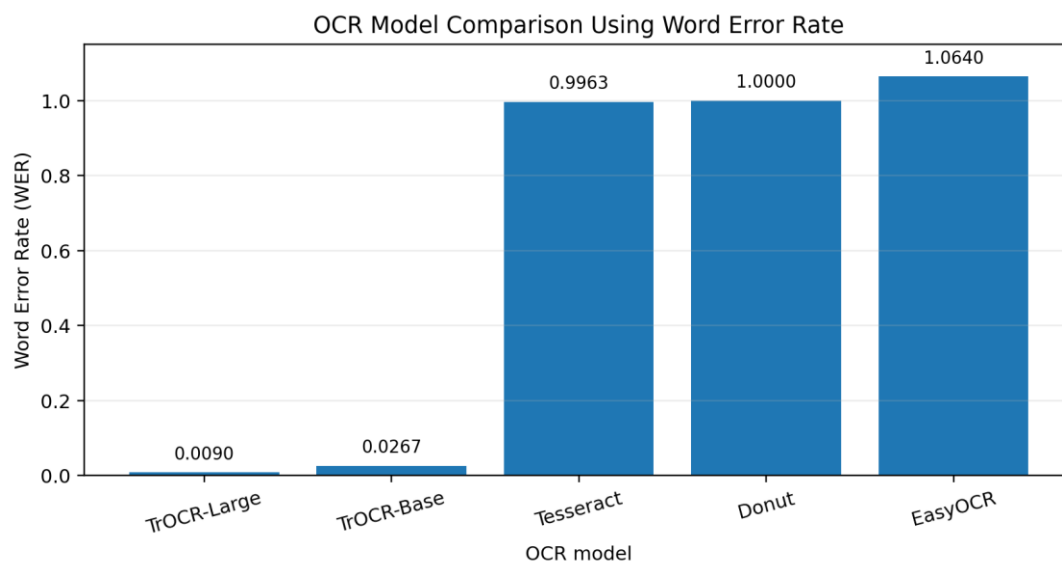


**Figure 7.** Recorded OCR model comparison using WER only; lower values indicate fewer word-level edit operations.

**Table 4.** Recorded OCR comparison under the project evaluation configuration.

| Model       | WER ↓  | Result within this configuration                                              |
|-------------|--------|-------------------------------------------------------------------------------|
| TrOCR-Large | 0.0090 | Lowest recorded word error.                                                   |
| TrOCR-Base  | 0.0267 | Low error with a smaller TrOCR variant.                                       |
| Tesseract   | 0.9963 | Most reference words require edit operations.                                 |
| Donut       | 1.0000 | One edit operation per reference word on average in this setup.               |
| EasyOCR     | 1.0640 | Insertions and other errors raise WER above 1, which is mathematically valid. |

Table 4 supports model selection for this prototype but does not establish a universal ranking across OCR tasks. Recognition results depend on language, image quality, segmentation, model fine-tuning, and evaluation protocol. The scientifically defensible conclusion is narrower: among the models evaluated in this configuration, the TrOCR variants produced the smallest word-level edit distance, and the large variant was selected for the integrated system.

## 5.3 | Error Sources and System-Level Interpretation

The largest observed performance change occurs when the task moves from isolated text lines to multi-line page images. This pattern is technically plausible because the page-level pipeline introduces several failure modes before recognition: a text region may be missed by the detector, adjacent lines may be merged, a crop may contain non-text material, or reading order may be reconstructed incorrectly. Each such event changes

the downstream token sequence and can create substitutions, deletions, or insertions even if the recognizer performs well on clean crops.

This interpretation also explains why reporting only a single “accuracy” value would conceal important information. WER provides a direct link between the pipeline and the sequence-edit errors that matter for downstream assessment processing. In practice, a low line-level WER supports transcription assistance, while the higher multi-line WER argues for teacher visibility of the recognized text before it contributes to a written-answer judgment.

## 5.4 | Scope, Validity, and Limitations

The experiment establishes technical feasibility for English handwriting and for the implemented document-processing configuration. It does not establish equivalent performance for other languages, very poor scans, mathematical notation, diagrams, or handwriting distributions that differ substantially from the evaluated data. The content-support module is also not assigned a quantitative educational-effect claim in this paper because the reported experiment provides recognition ground truth rather than a controlled learning-outcome study.

Potential construct threats are addressed by using WER as a sequence-recognition metric rather than a derived pseudo-accuracy. External validity remains bounded by the handwriting domain and document layouts. System validity is strengthened by documenting the detector model, thresholds, recognition model identifier, preprocessing sequence, and the distinction between line-level and page-level input. These boundaries keep the conclusions aligned with the evidence actually produced by the prototype.

## 6 | End-to-End Educational Workflows

### 6.1 | Teacher Authentication and Dashboard Access

The operational workflow begins with authenticated teacher access. After credential validation and role checking, the teacher is routed to a dashboard that exposes class management, assessment preparation, monitoring, and analytics functions. The navigation logic is summarized in Figure 8; the important design point is that the dashboard does not bypass server-side authorization when invoking protected functions.

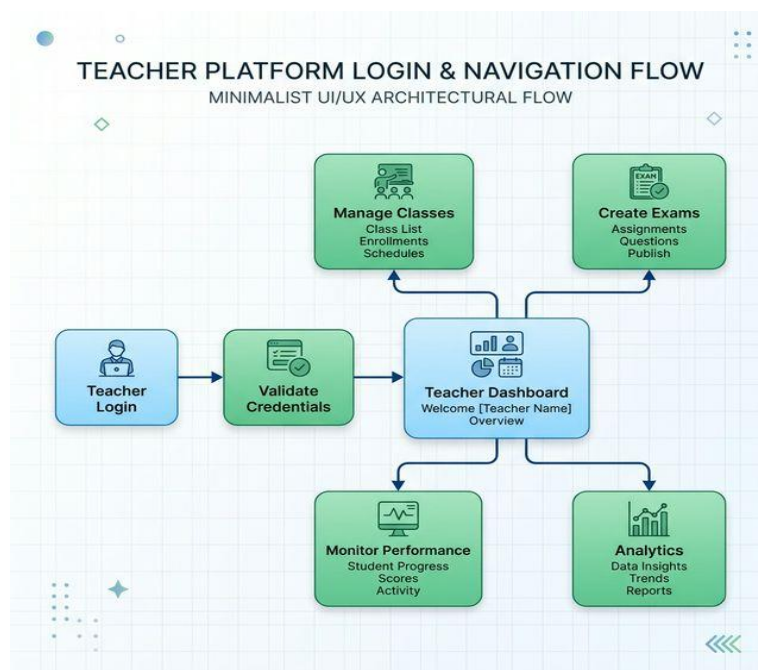


**Figure 8.** Teacher authentication and dashboard navigation workflow.

## 6.2 | Class Creation and Enrollment

From the dashboard, a teacher can create a class, define its academic context, and generate an enrollment mechanism for students. Figure 9 shows the sequence from class selection through class creation and student enrollment. The resulting membership record becomes the authorization context for lessons, examinations, attendance, and performance history.

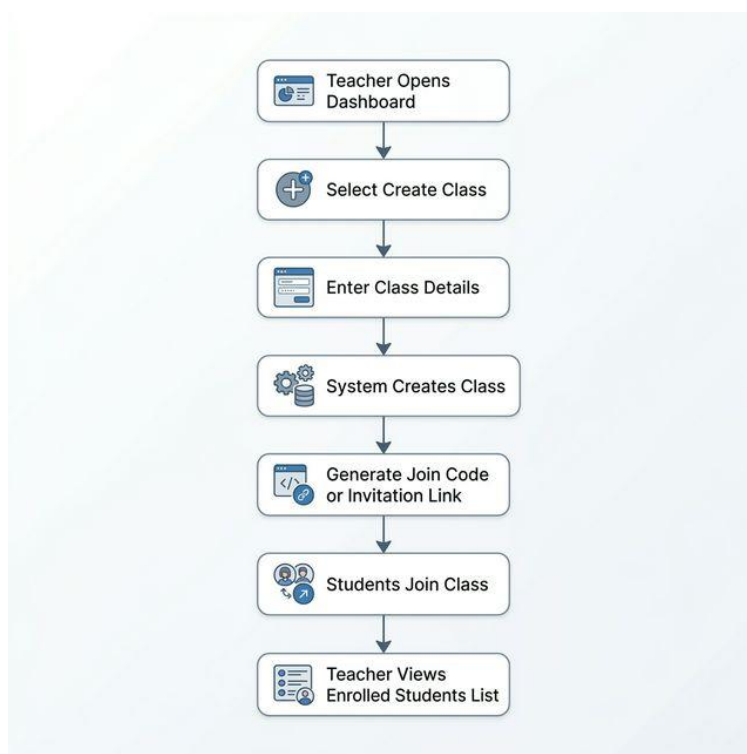


**Figure 9.** Class creation and student enrollment workflow.

## 6.3 | AI-Assisted Examination Preparation

Teachers can request draft questions by specifying the target class, topic, number of questions, difficulty level, and item type. The generated draft is not published automatically: the teacher can review, edit, replace, or regenerate items before release. Figure 10 illustrates this review loop. The workflow therefore uses generation to reduce preparation effort while preserving academic responsibility for the final assessment content.

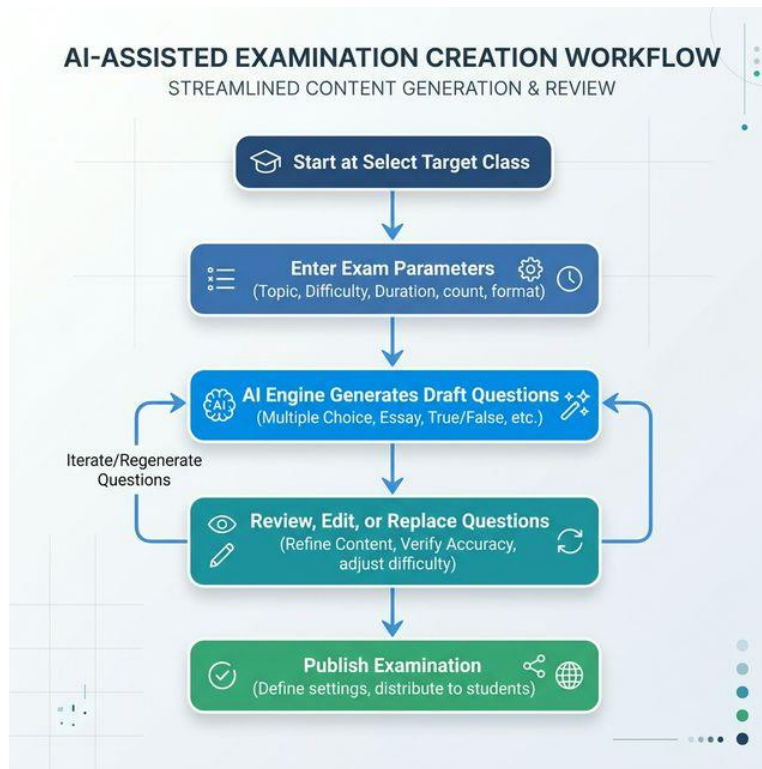


Figure 10. AI-assisted examination creation with explicit teacher review before publication.

## 6.4 | Timed Student Examination

When an examination becomes active, enrolled students can enter the attempt from their dashboard. Server-controlled timing governs the valid interval, while the client displays the remaining time. Figure 11 shows the attempt lifecycle, including manual submission and automatic submission when the valid period expires. This design reduces dependence on device-local time and makes examination state recoverable after a temporary interruption.

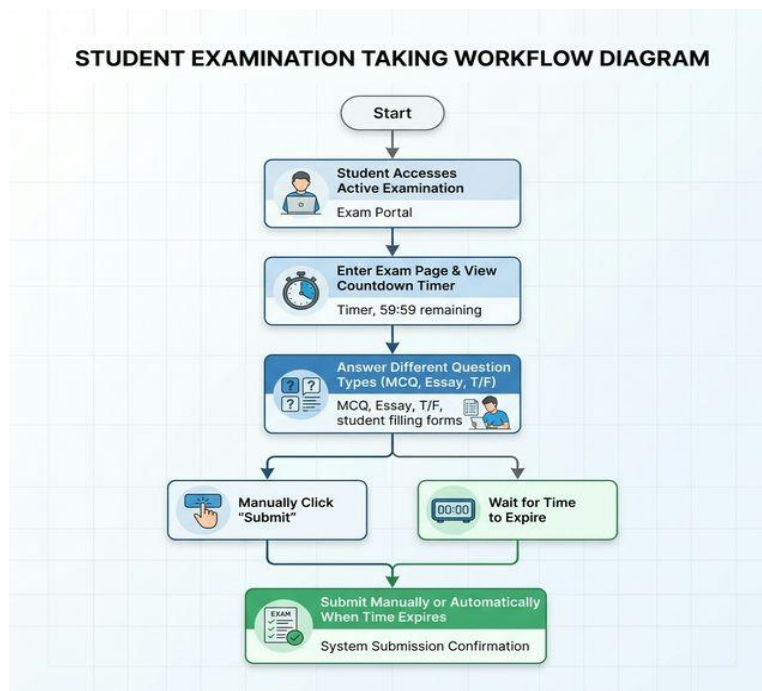


Figure 11. Student timed-examination workflow with server-controlled submission state.

## 6.5 | Result Verification and Feedback

After submission, the result can remain in a pending-review state while the teacher inspects written responses and any AI-produced transcription. Final marks are published only after verification. Figure 12 shows the transition from submission through review to student access. This sequence is central to the paper's human-in-the-loop contribution: automation accelerates processing, but the system maintains a distinct authorization step for the final academic result.

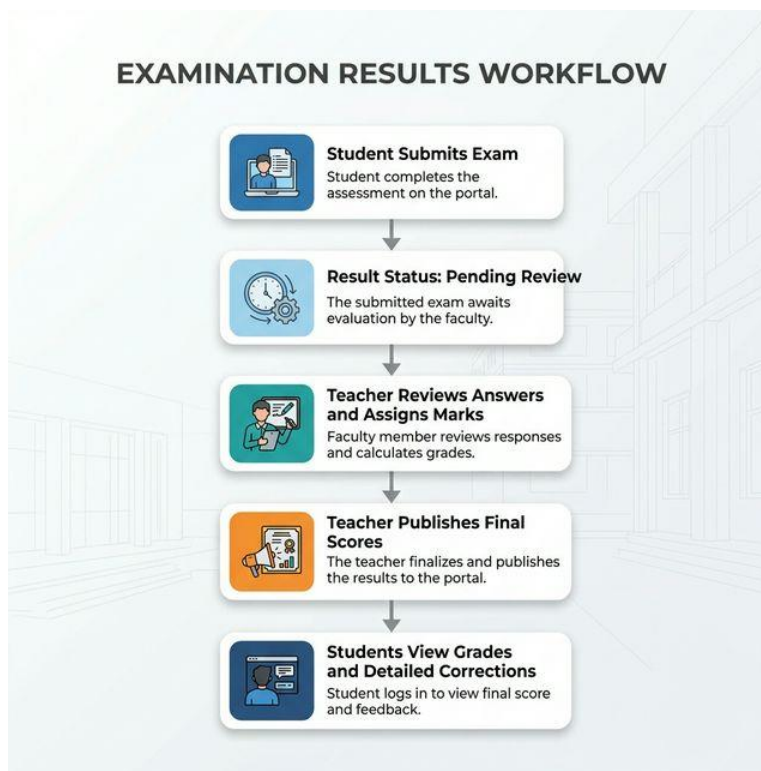


**Figure 12.** Examination-result workflow from submission to teacher verification and student feedback.

## 6.6 | Performance Support and Attendance

After verified results become available, Catalyst can aggregate repeated errors and weak topic areas to support follow-up recommendations such as revision topics, practice questions, or lesson suggestions. The same platform records attendance metadata for live sessions, including join and leave times where available. These features are presented as decision-support functions rather than evidence that learning outcomes have already improved; a controlled educational-impact study would be required to establish causal learning gains.

## 7 | Responsible Deployment and Educational Value

An educational AI platform processes data that may influence grades, progression, and learner self-perception. Responsible deployment therefore requires more than authentication. Catalyst's design uses data minimization at the workflow level, role-based access, explicit state transitions, teacher review for written responses, and separation between generated support content and final academic judgment. These principles align with UNESCO guidance that generative AI in education should be human-centred, privacy-aware, age-appropriate, and subject to meaningful oversight [15].

The practical value of the architecture lies in reducing repeated administrative operations while preserving auditability. Teachers obtain a single workflow for class management, assessment creation, document transcription, result review, and follow-up. Students obtain a consistent environment for examinations, feedback, progress visibility, and targeted support. For institutions, the modular backend makes it possible to

replace or retrain individual AI components without redesigning the complete application. This separation is particularly important because educational AI models evolve faster than the academic processes they must serve.

The paper does not treat automation as synonymous with educational quality. Recent AIED and human-centred analytics research emphasizes that technical systems must be evaluated in relation to pedagogy, agency, reliability, and the actions they enable [1,5–7]. Catalyst’s contribution is therefore best understood as an integration architecture with quantified handwriting performance and explicit control boundaries, providing a foundation for subsequent classroom studies of usability, workload reduction, feedback quality, and learning outcomes.

## 8 | Conclusion and Future Work

This paper presented Catalyst as an integrated AI-enabled educational platform that connects handwritten document processing with assessment workflow automation, teacher verification, content support, and performance-oriented follow-up. The principal technical contribution is a two-stage document pipeline in which DB-based text detection isolates line regions and TrOCR converts the crops into text, while the broader system contribution is the separation of machine assistance from final academic authorization. This distinction turns OCR from an isolated benchmark component into part of a controlled educational process.

The quantitative evaluation shows  $WER = 0.0090$  for isolated handwritten lines and  $WER = 0.1925$  for multi-line document processing. The model comparison further shows that the TrOCR variants produced the lowest recorded WER in the project configuration. Replacing the previous WER-to-accuracy transformation with direct WER reporting also removes a methodological ambiguity and correctly accommodates cases in which insertion errors produce WER values above 1.

Future work should target the part of the pipeline that the current evidence identifies as most vulnerable: page-level detection, segmentation, and reading-order reconstruction. Additional experiments should extend recognition to multilingual handwriting, mathematical notation, and more diverse document layouts. The educational layer should be evaluated separately through controlled studies of teacher workload, feedback usefulness, learner engagement, and calibration of performance recommendations. For generative content support, future releases should incorporate grounding, provenance, and teacher-configurable constraints before generated material is used in high-stakes contexts. These extensions preserve the central design principle of Catalyst: AI should increase the speed and usefulness of educational workflows while keeping evidence, control, and final academic responsibility visible to human users.

## Author Contribution

All authors contributed equally to this study.

## Funding

This research received no external funding.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

- [1] Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252, 124167. <https://doi.org/10.1016/j.eswa.2024.124167>

- [2] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16, 39. <https://doi.org/10.1186/s41239-019-0171-0>
- [3] Al-Fraihat, D., Joy, M., Masa'deh, R. E., & Sinclair, J. (2020). Evaluating e-learning systems success: An empirical study. *Computers in Human Behavior*, 102, 67–86. <https://doi.org/10.1016/j.chb.2019.08.004>
- [4] Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C., & Srinivasa, A. R. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 6, 100206. <https://doi.org/10.1016/j.caeai.2024.100206>
- [5] Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6, 100215. <https://doi.org/10.1016/j.caeai.2024.100215>
- [6] Paulsen, L., & Lindsay, E. (2024). Learning analytics dashboards are increasingly becoming about learning and not just analytics—A systematic review. *Education and Information Technologies*, 29, 14279–14308. <https://doi.org/10.1007/s10639-023-12401-4>
- [7] Graf, L., Bassner, P., Anzinger, M., Dietrich, F. T. J., Krusche, S., & Poquet, O. (2026). Interactive learning dashboards: Rethinking learning visualisations as engagement tools. *Education and Information Technologies*. <https://doi.org/10.1007/s10639-026-14082-1>
- [8] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florêncio, D. A. F., Zhang, C., Li, Z., & Wei, F. (2023). TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11), 13094–13102. <https://doi.org/10.1609/aaai.v37i11.26538>
- [9] Marti, U.-V., & Bunke, H. (2002). The IAM-database: An English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1), 39–46. <https://doi.org/10.1007/s100320200071>
- [10] Liao, M., Wan, Z., Yao, C., Chen, K., & Bai, X. (2020). Real-time scene text detection with differentiable binarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 11474–11481. <https://doi.org/10.1609/aaai.v34i07.6812>
- [11] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
- [12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
- [13] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2022). OCR-free document understanding transformer. In *Computer Vision – ECCV 2022* (pp. 498–517). Springer. [https://doi.org/10.1007/978-3-031-19815-1\\_29](https://doi.org/10.1007/978-3-031-19815-1_29)
- [14] Smith, R. (2007). An overview of the Tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (pp. 629–633). IEEE. <https://doi.org/10.1109/ICDAR.2007.4376991>
- [15] Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO.
- [16] Memarian, B., & Doleck, T. (2024). A review of assessment for learning with artificial intelligence. *Computers in Human Behavior: Artificial Humans*, 2(1), 100040. <https://doi.org/10.1016/j.chbah.2023.100040>

**Disclaimer/Publisher's Note:** The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.