Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE
Paper ID #45749

Automating Structured Information Extraction from Images of Academic Transcripts Using Machine Learning

Declan Kirk Bracken, University of Toronto

Declan Bracken is an M.Eng. student at the University of Toronto in the department of Mechanical and Industrial Engineering pursuing an emphasis in Analytics. This paper is the final product of an 8 month M.Eng. project supervised by Professor Sinisa Colic and it's work is intended for implementation into the admissions process at the University of Toronto's M.I.E department.

Dr. Sinisa Colic Ph.D., University of Toronto

Dr. Colic is an Assistant Professor, Teaching Stream with the Department of Mechanical and Industrial Engineering. He completed his PhD at the University of Toronto in the area of personalized treatment options for epilepsy using advanced signal processing techniques and machine learning. Dr. Colic currently teaches several courses at University of Toronto covering a broad range of topics in mechatronics, data science and machine learning / deep learning.

Automating Structured Information Extraction from Academic Transcript Images Using Machine Learning

Abstract

The admission process for post-secondary institutions requires staff to spend countless hours manually reviewing applications and student transcripts to make critical decisions about their academic future. Academic transcript images are tedious to read and transcribe due to their myriads of visual features, such as colored backgrounds, watermarks, multicolumn layouts, and small text. To streamline this process, this report investigates the development of an AI system specifically designed for transcribing grade data from images of academic transcripts into organized tables. While models for table extraction are not novel, existing methods are limited when dealing with academic transcripts due to their unique features and a lack of representation in preexisting datasets used for training. To our knowledge, this report implements the first labeled open-source dataset of purely academic transcript images used for training computer vision-based machine learning algorithms. Two primary approaches for image-to-text table reconstruction were explored; the first is a pipeline comprising a YOLOv8 object detection model, a Tesseract OCR engine, and a Mistral7b large language model (LLM). The second option implemented a fine-tuned multimodal language model (MiniCPM-Llama3-V-2_5). The multimodal LLM showed superior accuracy on a small test set, able to locate, read, and reorganize noisy tabular data into organized strings. Future work could greatly improve on this solution by expanding the dataset with a greater diversity of images and experimenting with different models or training methods. This work provides a platform with which admissions data may be used to predict student success and better track student progress over their academic career through data driven analysis.

Introduction

The graduate admissions committee at a top-ranked university reviews over 1,000 applicants annually and is a cornerstone of academic excellence, but the admissions process remains labor-intensive. Staff are required to manually review numerous student documents, particularly academic transcripts, which contain essential data such as grades and course credits that must be meticulously analyzed to ensure fair and consistent decisions. Since transcripts hail from a variety of institutions globally, each with different formatting nuances as well as curriculum, difficulties arise when comparing student performance. This manual effort slows the admissions process and places a significant burden on staff.

A system which transcribes and organizes transcript data into structured formats, particularly grade tables, could drastically improve the efficiency and consistency of admissions decisions. Such a tool could enable downstream applications, such as analytical dashboards or predictive models of student success, by creating a foundational database of student grade data. Existing AI solutions for table extraction from images are, however, ill-suited to the unique demands of academic transcripts.

Optical Character Recognition (OCR) is a computational system dedicated to extracting text from

pure images into a machine-readable format [1]. While OCR technology is mature and widely used, challenges persist in detecting and reconstructing the format of complex tables, particularly in specialized documents like academic transcripts. Recent advancements in computer vision for document analysis, including TableNet, CascadeTabNet, and LayoutLM, have greatly improved table detection and structure recognition tasks. TableNet, designed specifically for table detection in document images, has shown promise in recognizing tables from structured documents like financial statements [2]. While its performance on structured tables is strong, it faces challenges when dealing with academic transcripts, which often feature a mix of tabular and non-tabular content, making traditional layout models less effective. CascadeTabNet and LayoutLM, more recent models, have also extended the capabilities of document structure recognition by incorporating multi-stage detection and leveraging transformer-based approaches for richer semantic understanding [3] [4]. However, these models, while effective on datasets like ICDAR, often struggle with the unique challenges of academic transcripts as they mainly feature scientific journals, handwritten notes, financial statements, and other documents which differ greatly in visual structure [5].

Academic transcripts vary significantly between institutions, featuring irregular or inconsistent table structures and noise introduced by watermarks and background patterns. Key features such as multicolumn formats, inconsistent text spacing, and variably placed or even missing headers further complicate the extraction and reconstruction tasks. Moreover, the absence of cleanly delineated table borders and the reliance on empty space to separate data further hinder the application of existing table recognition models. To our knowledge, there are no publicly available datasets of academic transcripts which can be used to train these models, likely due to the sensitive nature of the data and legal restrictions on sharing student information.

This project addresses these gaps by curating and labeling a dataset of publicly available academic transcript images and developing a privacy-compliant AI system that can be hosted locally to extract and organize grade data into machine-readable tables.

Constraints

The development of an AI system for transcript analysis is governed by key constraints focused on privacy, computational requirements, data availability, and budget limitations. Given the sensitivity of student transcript data, the system must operate entirely on a local or university-owned machine to align with Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) [6]. This ensures strict control over data handling, minimizing risks associated with internet transmission and unauthorized access. Additionally, the system must meet reasonable computational requirements: for personal machines, modern multi-core CPUs, 16–32 GB of RAM, SSD storage, and mid-range GPUs are sufficient. Finally, the project is self-funded, requiring creative resource optimization to navigate limited financial support. This includes leveraging open-source tools, sparing use of low-cost cloud computing, and manual data labeling where necessary, ensuring a high-quality, scalable solution within these constraints.

Dataset Image Collection

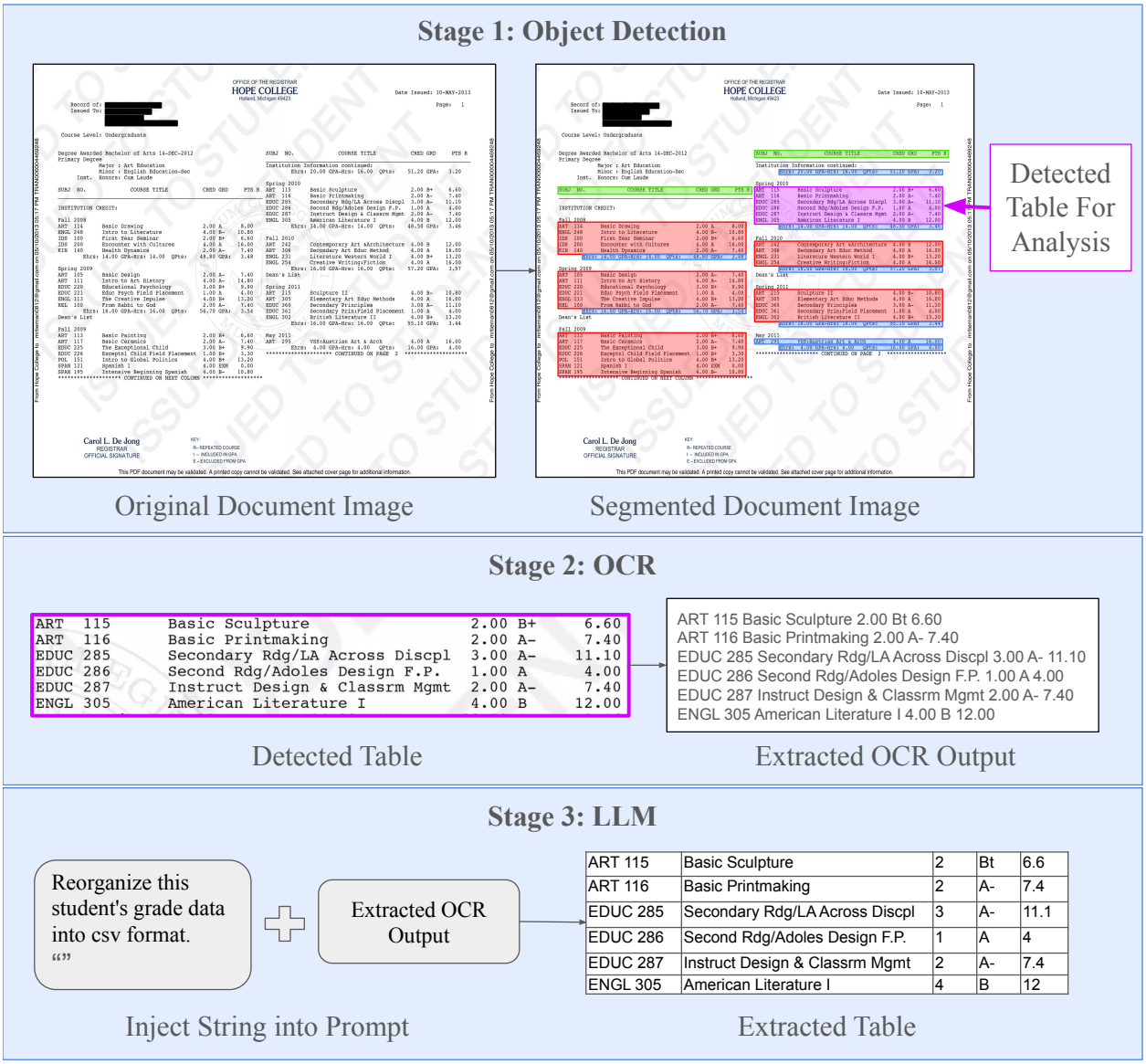
A small dataset of 200 transcript images was collected, primarily sourced from the website SlideShare due to its abundance of high-quality and publicly accessible transcript images [7]. After cleaning for duplicates and irrelevant pages, the dataset was refined to 126 images. The images exhibit significant variability in resolution, loosely classified into three groups: small images (640×640 pixels), sourced mainly from Google, and larger images of varying aspect ratios (2048×1582 and 2048×2700 pixels) from SlideShare. The higher-resolution images, which are representative of the data typically encountered by admissions staff, make up the majority of the dataset and ensure its relevance for training and evaluation.

Optical Character Recognition

When designing any system that requires text retrieval from an image, an obvious starting point is OCR. There are several open-source OCR engines available, and for this project Tesseract was chosen for its accuracy, ease of use, language-specific customizations, and fine-tuning capability [8]. Tesseract is capable of performing inference with a variety of different configurations, such as reading an image as a paragraph, line, word, or as sparse text, to name a few. Tesseract, however, struggles with images containing complex layouts, such as multiple text columns or blocks, as it cannot accurately determine the logical reading order and instead tries to interpret the content as a single, continuous paragraph of text. Full-page OCR processing of a transcript produces a mixed output of paragraphs, headings, and tables, making it difficult to extract clean and organized information. Therefore, to ensure that the output is accurately formatted, it is crucial to filter out unnecessary information and segment key content into individual, easily readable blocks before applying OCR. For example, consider a table which has been isolated from the document and is passed as an image to OCR. The resulting string may accurately retrieve the contents of the table in ordered rows, however, the location of column breaks will still be lost. We therefore also need a method of preserving the column structure of the table after applying OCR.

Computer Vision

To address these OCR limitations, we implemented a three-stage pipeline comprising an object detection model (YOLOv8), optical character recognition (Tesseract), and a large language model (Mistral7b) [9] [10]. The YOLOv8 model preprocesses student transcripts by segmenting and cropping grade data, while the LLM post-processes the OCR output by reorganizing it into comma-separated-value (CSV) format. This is a novel approach which relies on the LLM to understand the semantic relationships between words in the tables and intuitively determine where column separations must be placed. The three-stage pipeline is illustrated with an example transcript in Figure 1. In stage 1, the document is passed to the YOLOv8 model which segments the transcript into regions of interest. Each region is cropped by their predicted bounding boxes and individually passed to OCR (stage 2). Each string of grade data is then concatenated into a single piece of text (not pictured in Figure 1), which is then given to a Mistral7b model with a prompt asking the LLM to reorganize the text into CSV format. In Figure 1 the final output is presented as a table instead of a CSV formatted string for clarity.



ART 115	Basic Sculpture	2	Bt	6.6
ART 116	Basic Printmaking	2	A-	7.4
EDUC 285	Secondary Rdg/LA Across Discpl	3	A-	11.1
EDUC 286	Second Rdg/Adoles Design F.P.	1	A	4
EDUC 287	Instruct Design & Classrm Mgmt	2	A-	7.4
ENGL 305	American Literature I	4	B	12

Figure 1: Three stage pipeline. The purple grade table in the segmented document image of Stage 1 is used as a sample input to illustrate the remainder of the pipeline.

Of the three pipeline stages, only the vision component must be trained on a custom transcript-based dataset. The YOLOv8 computer vision model was trained using the transcript images collected in Section 2. Multiple labeled object detection datasets were created which tested different labeling philosophies and how information should be segmented. The final labeling schema features three object types: Grade tables, single-row tables, and headers. These three classes are visible in the Segmented Document Image from Stage 1 of Figure 1 as the red, green, and blue bounding boxes respectively. Grade tables form the bulk of detections and are the multi-row tables whose data we are attempting to extract. The tables' headers are detected as a separate class from the grade table due to the prevalence of certain transcript formats with headers offset vertically from the data, and hence need to be cropped separately. The single-row table class was

created as a method of improving the accuracy of the model on detecting individual lines. The vision model cannot read and understand text semantics, only visual patterns, hence tables which featured only a single row of data originally had a very low detection rate during early testing as they were difficult to distinguish from other pieces of single line text. To compensate for this edge case, the single-row table class was created for any relevant columnar data whose structure is not repeated for 2 or more consecutive rows.

Multi-Modal LLM

As an alternative to the three-stage pipeline, implementation of an open-source multi-modal LLM (MM-LLM) was also examined. Implementation of a single model which is capable of both seeing the layout of a document and understanding the semantic meaning of it's text is intuitively more flexible and powerful than a pipeline with distinct feature engineering and extraction steps. Only one dataset is required for training, whereas the pipeline would require multiple datasets, one for each distinct model, to iteratively improve accuracy. To validate this idea, this project implemented MiniCPM-Llama3-V-2_5, an open-source LLM built off Meta's Llama3 model [11]. With only 8.5 billion parameters, this model provides a robust solution for reading images, even outperform GPT-4V-1106 on OCRBench, but with significantly lower computational demands [11]. For training, an image-text pair dataset was created by manually transcribing the grade data from each sample transcript into CSV format. This image-text pair dataset would also serve as a means of testing the accuracy for the end-to-end pipeline, since the final output of either method is a string in CSV format.

Evaluation Methods

While both of the examined approaches for table reconstruction require some level of computer vision and text localization, assessment of overall performance is limited to string comparisons due to the nature of the final output as a string. The evaluation of table reconstruction focuses on three complementary metrics: Levenshtein Distance, Token F1 Score, and ROUGE-L Score. Levenshtein Distance measures the character-level similarity between the model output and ground truth, identifying minor text extraction errors [12]. Token F1 Score balances precision and recall to evaluate content accuracy at the token level, ensuring the correct identification of table elements [13]. ROUGE-L Score assesses structural fidelity by capturing the longest common subsequence, which reflects the preservation of order in tabular data [14]. Together, these metrics comprehensively evaluate the accuracy of reconstructed tables, addressing both content correctness and structural alignment. From the original dataset of 126 images, 12 were randomly selected to create the unseen test set.

Results

As shown in Table 7, the modular pipeline combining vision, OCR, and LLM components achieved an average Levenshtein Distance of 859.33, indicating fewer character-level discrepancies compared to the multimodal LLM which had an average distance of 924.42. However, the MM-LLM demonstrated higher performance in Token F1 Score and ROUGE-L Score, achieving 0.3186 and 0.3612, respectively, compared to the modular pipeline's 0.1962 and 0.2162. These results gener-

ally indicate a high degree of error in predictions.

Metric	Modular Pipeline	MM-LLM
Levenshtein Distance	859.33	924.42
Token F1 Score	0.1962	0.3186
ROUGE-L Score	0.2162	0.3612

Table 1: Table reconstruction performance metrics.

Investigating results on individual transcripts reveals drastic changes in performance between samples from the test set. The MM-LLM scored as high as 0.91 and 0.93 on Token-F1 and ROUGE-L respectively, but also as low as 0.04 and 0.10. The pipeline did not score as highly, but still showed variance between samples, with a max Token-F1 score of 0.51. To illustrate model performance further, a sample from the test set is presented in Figure 2, with the respective model outputs converted from csv strings to tables in Figures 3 and 4.

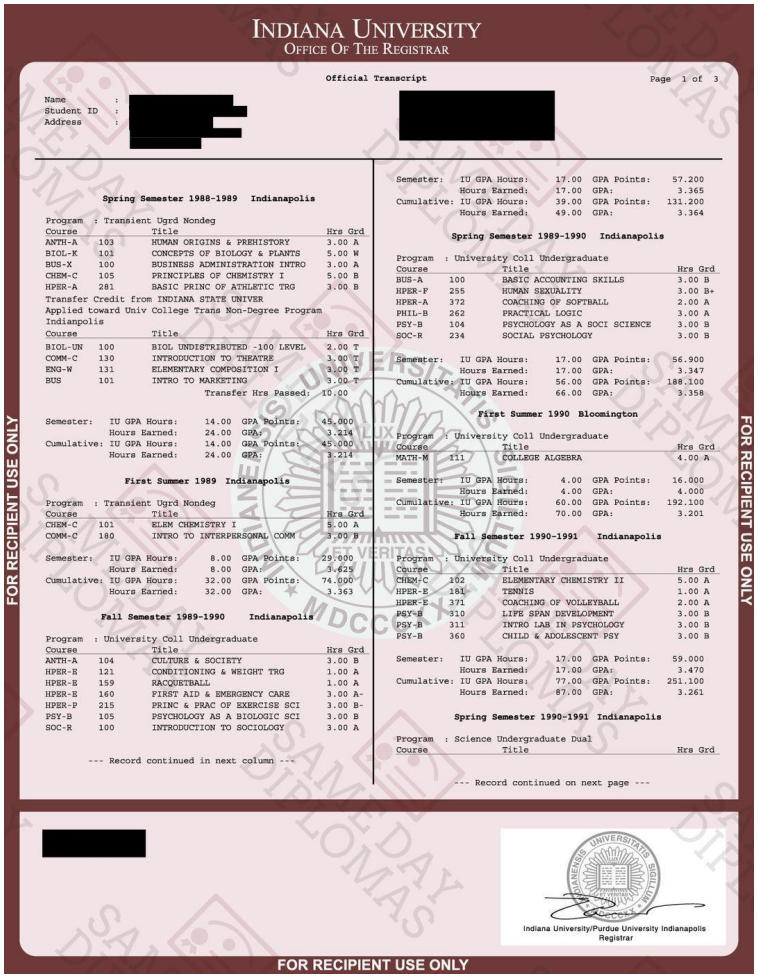


Figure 2: Sample transcript from the test set. Inference results can be seen in Figures 3 and 4.

Comparing the table reconstructions from Figures 3 and 4, neither method completely captures all courses from the original transcript image, but still manages to reconstruct most of the orig-

Figure 3: Pipeline prediction.

Course Title	Hrs	Grd.
ANTH-A 104 CULTURE & SOCIETY	3.00	B
HPER-E 121 CONDITIONING & WEIGHT TRG	1.00	A
HPER-E 159 RACQUETBALL	1.00	A
HPER-E 160 PIRST AID & EMERGENCY CARE	3.00	A-
HPER-P 215 PRINC & PRAC OF EXERCISE SCI	3.00	B-
PSY-B 105 PSYCHOLOGY AS A BIOLOGIC SCI	3.00	B
SOC-R 100 INTRODUCTION TO SOCIOLOGY	3.00	A
BUS-A 100 BASIC ACCOUNTING SKILLS	3.00	B
HPER-F 255 HUMAN SEXUALITY	3.00	Bt
HPER-A 372 COACHING OF SOFTBALL	2.00	0
PHIL-B 262 PRACTICAL LOGIC	3.00	A
PSY-B 104 PSYCHOLOGY AS A SOCI SCIENCE	3.00	B
SOC-R 234 SOCIAL PSYCHOLOGY	3.00	B
ANTH-A 103 HUMAN ORIGINS & PREHISTORY	3.00	A
BIOL-K 101 CONCEPTS OF BIOLOGY & PLANTS	5.00	W.
BUS-X 100 BUSINESS ADMINISTRATION INTRO	3.00	A.
CHEM-C 105 PRINCIPLES OF CHEMISTRY I	5.00	B
HPER-A 281 BASIC PRINC OF ATHLETIC TRG	3.00	B
BIOL-UN 100 BIOL UNDISTRIBUTED -100 LEVEL	2.00	T
COMM-C 130 INTRODUCTION TO THEATRE	2. g0%7)	

Figure 4: MM-LLM prediction.

Course Title	Hours Earned	Grade
ANTH-A 103 HUMAN PRINCIPLES & HISTORY	3.00	A
BIOL-K 101 CONCEPTS OF BIOLOGY & PLANTS	5.00	W
BUS-X 100 BUSINESS ADMINISTRATION INTRO	3.00	C
CHEM-C 105 PRINCIPLES OF CHEMISTRY I	5.00	B
HPER-A 281 BASIC PRINCE OF ATHLETIC TRG	3.00	B
BIOL-UNI 100 INTRO TO BIOLOGY UND	2.00	T
COMM-C 130 INTRODUCTION TO THEATER	3.00	T
ENG-W 131 ELEMENTARY COMPOSITION I	3.00	T
BUS 101 INTRO TO MARKETING	3.00	T
CHEM-C 101 ELEMENTARY CHEMISTRY I	3.00	B
COMM-C 180 INTRO TO INTERPERSONAL COMM	3.00	B
ANTH-I 104 CULTURE & SOCIETY	3.00	B
HPER-E 121 CONDITIONING & WEIGHT TRG	1.00	A
HPER-E 159 RACQUETBALL	1.00	A
HPER-E 160 FIRST AID & EMERGENCY CARE	3.00	A-
HPER-P 215 PRIN & PRAC OF EXERCISE SCI	3.00	B-
PSY-B 105 PSYCHOLOGY AS A BIOLOGIC SCI	3.00	B
SOC-R 100 INTRODUCTION TO SOCIOLOGY	3.00	A

inal text in an acceptable columnar format. The 'course' and 'title' columns are merged, but no information has been lost. For the MM-LLM, the original course information and grade data is near-perfectly reconstructed, even maintaining the chronological order the courses were taken in, but only for the left half of the page. Missing the right column's data, the MM-LLM achieves a Levenshtein Distance, Token-F1 score, and ROUGE-L score of 575, 0.46, and 0.50 respectively. In comparison, the pipeline's output recalls a similar number of courses, but from various different sections of the page. The pipeline also has some artifacts from the OCR inference, such as periods following grades, or plus symbols interpreted as the letter 't'. The last line of the pipeline features a malformed string in the 'Hrs' column, likely due to the bounding box from the vision model cropping some of the characters in the table, resulting in further errors during OCR. The pipeline achieves a relatively similar Levenshtein Distance, Token-F1 score, and ROUGE-L score of 738, 0.51, and 0.56 respectively on this transcript.

Discussion

In assessing the performance of our two approaches for transcript data extraction, we used string comparison metrics like Levenshtein Distance, Token F1 Score, and ROUGE-L Score. While these metrics are useful, they primarily focus on character or token-level accuracy, which may not fully capture the complexities of multimodal tasks. A multimodal approach involves multiple stages of data extraction including vision, OCR, and language models—which can result in discrepancies that purely string-based comparisons fail to address. Minor variations in phrasing or formatting could distort results, even if the meaning remains intact. A more appropriate assessment method could involve evaluating the semantic accuracy of outputs, using human-in-the-loop validation or task-specific metrics that account for both content accuracy and contextual relevance.

The average results from Section 7 indicate a lack of accuracy in table reconstruction between either the pipeline or MM-LLM methods. Examining individual test cases, however, the MM-LLM demonstrated promising results on some specific samples, achieving a Levenshtein Dis-

tance as low as 38.0 and high Token-F1 (0.91) and ROUGE-L (0.93) scores, indicating accurate character and token-level predictions. Analysis of the MM-LLM's performance on the sample transcript from Figure 2 shows a clear ability to read and organize long sequences of text. These results suggest that the MM-LLM has the potential to perform well, even on complex or noisy images, but may require a combination of systematic changes to the dataset, model, or preprocessing techniques in order to be viable. While the pipeline clearly has a capacity to succeed on some transcripts, issues in the YOLOv8 model cropping out courses, the OCR engine misidentifying characters, and the LLM's imperfect recall makes the system more difficult to iterate on. Since the results for both methods were obtained from a dataset with fewer than 150 samples, increasing the dataset size is a straightforward way to potentially improve performance.

The implementation of either the MM-LLM or pipeline for automatic transcript data extraction faces other challenges in interpretability, privacy, and computational requirements. Interpretability is a key issue, as these models often function as "black boxes", making it difficult to explain predictions and diagnose issues. During development and testing, missing courses and even model hallucinations were common artifacts in the output strings, often comprising course names and grades taken from different parts of the transcript and paired together. Occasionally, and especially for the pipeline system, poor input data, such as mistakes in OCR, could result in the LLM making up new courses and grades entirely. Privacy is also a concern, particularly with sensitive student data, as compliance with regulations like PIPEDA is essential. Additionally, the computational demands of LLMs require powerful hardware, which can be cost-prohibitive for institutions with limited resources. Balancing these challenges is crucial for effective deployment in educational contexts.

If the MM-LLM is improved and implemented, automatic transcript data extraction could provide valuable insights for tracking student performance. By accurately reconstructing transcript data, it could enable a deeper analysis of individual student progress over time, identifying patterns in grades, course completion, and skill development. For instance, if the system could track course-specific performance trends, it could highlight areas where a student consistently struggles, helping educators pinpoint subjects or skills that require additional focus. This could lead to more personalized learning interventions and better-targeted academic support, ultimately enhancing student outcomes.

Conclusion

This paper explored the development of a new dataset of student academic transcripts and an accompanying AI system designed to transcribe and structure grade data into organized tables. Two approaches were examined: a modular pipeline consisting of an object detection model, OCR engine, and LLM, as well as a fine-tuned multimodal LLM. While both models exhibited difficulty in complete table reconstruction, they also demonstrated some promising initial results. Given the small dataset size and limited computational resources, the models were able to locate tables, accurately read small, noisy text, and reorganize large sequences of strings into a columnar structure. These results suggest that with more data and continuous improvement, these systems could be implemented to greatly support the admissions process in the future.

References

[1] R. Avyodri, S. Lukas, and H. Tjahyadi, "Optical character recognition (ocr) for text recognition and its post-processing method: A literature review," in *2022 1st International Conference on Technology Innovation and Its Applications (ICTIIA)*, 2022, pp. 1–6.

[2] S. Paliwal, V. D, R. Rahul, M. Sharma, and L. Vig, "Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images," *CoRR*, vol. abs/2001.01469, 2020. [Online]. Available: http://arxiv.org/abs/2001.01469

[3] D. Prasad, A. Gadpal, K. Kapadni, M. Visave, and K. Sultanpure, "Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents," *CoRR*, vol. abs/2004.12629, 2020. [Online]. Available: https://arxiv.org/abs/2004.12629

[4] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "Layoutlm: Pre-training of text and layout for document image understanding," *CoRR*, vol. abs/1912.13318, 2019. [Online]. Available: http://arxiv.org/abs/1912.13318

[5] *Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part III*. Berlin, Heidelberg: Springer-Verlag, 2023.

[6] "Personal information protection and electronic documents act, sc 2000, c 5," https://canlii.ca/t/56ck6, 2000, retrieved on 2025-01-13.

[7] Slideshare.net, "Slideshare.net," https://www.slideshare.net/, accessed Jan. 13, 2025.

[8] R. Smith, "An overview of the tesseract ocr engine," in *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 629–633. [Online]. Available: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/33418.pdf

[9] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time flying object detection with yolov8," 2024. [Online]. Available: https://arxiv.org/abs/2305.09972

[10] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[11] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, Q. Chen, H. Zhou, Z. Zou, H. Zhang, S. Hu, Z. Zheng, J. Zhou, J. Cai, X. Han, G. Zeng, D. Li, Z. Liu, and M. Sun, "Minicpm-v: A gpt-4v level mllm on your phone," 2024. [Online]. Available: https://arxiv.org/abs/2408.01800

[12] R. Haldar and D. Mukhopadhyay, "Levenshtein distance technique in dictionary lookup methods: An improved approach," 2011. [Online]. Available: https://arxiv.org/abs/1101.1232

[13] L. Wang, Y. Shen, S. Peng, S. Zhang, X. Xiao, H. Liu, H. Tang, Y. Chen, H. Wu, and H. Wang, "A fine-grained interpretability evaluation benchmark for neural nlp," 2022. [Online]. Available: https://arxiv.org/abs/2205.11097

[14] K. Ganesan, "Rouge 2.0: Updated and improved measures for evaluation of summarization tasks," 2018. [Online]. Available: https://arxiv.org/abs/1803.01937