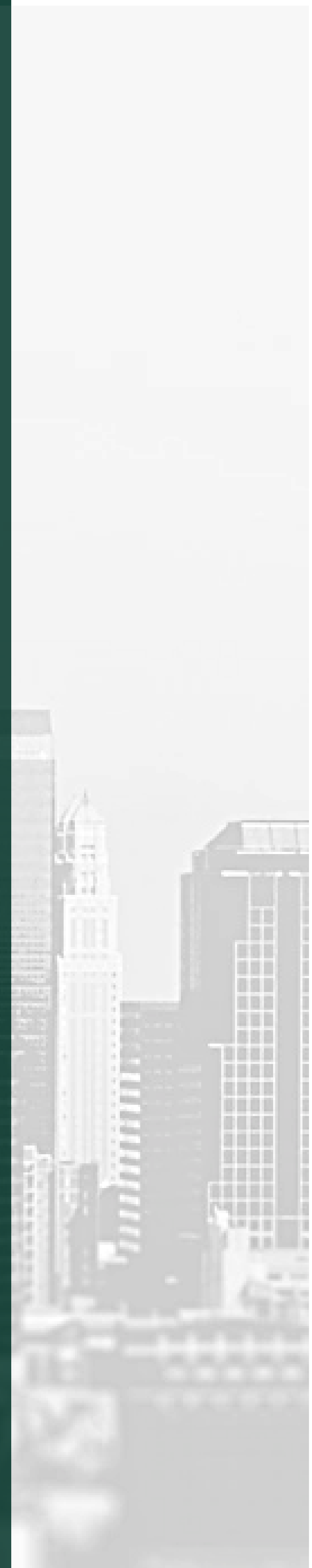
MICHIGAN STATE UNIVERSITY

UNIVERSITY OUTREACH AND ENGAGEMENT

INNOVATE MICHIGAN!

2026 Student-Led, Faculty-Guided Technical Assistance Project Report

Digitizing and Analyzing Timber Lease Financial Records for Economic Development



WRITTEN BY:

Bibin Varghese (MS Data Science) with Karen Sandy, Christine Yoo, Tianjian Xing, Corei Laury

LUCE COUNTY TIMBER FUND

Financial Data Pipeline & Analysis

Final Report - July 2026

Community-Engaged Data Project with Luce County, Michigan

Prepared by	Bibin Varghese (MS Data Science) with Karen Sandy, Christine Yoo, Tianjian Xing, Corei Laury
Submitted to	Center for Community and Economic Development (CCED), Michigan State University
Community Partner	Tammy Lasley-Henry, Executive Director at Luce County Economic Development Corporation
Course	CMSE 495 — Experiential Learning in Data Science
Repository	gitlab.msu.edu/xingtia2/luce-lumber (private)
NDA Status	All processing performed locally; no county data transmitted externally

Note: *This project is supported in part pursuant to the receipt of financial assistance from the MSU Center for Community and Economic Development's Regional Economic Innovation. The opinions expressed in the statements, findings, conclusions, and recommendations are the authors' sole responsibility and do not necessarily reflect the views of Michigan State University.*

Abstract

This project partners Michigan State University graduate students with Luce County, Michigan to digitize and analyze 40 years of financial records for the county's timber lease fund (1983–2025). The records come from two distinct sources: scanned PDF documents (handwritten invoices, registers, and statements from 1983–2022) and structured QuickBooks digital exports (2014–2025). The overarching goal is to produce a reliable yearly summary of money flowing into and out of the timber fund, and to deliver tools the county can operate independently.

This report documents two data pipelines developed by the graduate team. The first uses Optical Character Recognition (OCR) combined with a locally running AI model (Phi-3 via Ollama) to extract transactions from scanned historical documents in a fully NDA-compliant manner. The second analyzes the QuickBooks export to categorize 351 transactions and generate yearly financial summaries. An interactive local dashboard built with Reflex visualizes the QuickBooks findings.

Testing on a real 1990 document (1990_002651.pdf, 20 pages) revealed both the potential and the limitations of the pipeline: 39 of 41 transactions were successfully extracted, with 61% classified at high confidence. Key limitations include OCR struggles with handwriting and mixed document types (e.g., checks, journal entries, vouchers), and AI struggling to extract rows of dense data and occasionally fabricating plausible values. These findings directly inform what the county should realistically expect when adopting this system.

1. Background and Motivation

Luce County, Michigan manages a multi-thousand-acre timber lease program on state-owned land. Timber companies pay the county for the right to harvest trees, and that revenue flows into a revolving loan fund (RLF) that supports local economic development. As the county prepares for future lease renewals with the State of Michigan, it needs a clear, documented history of how these funds have been received and spent over the past four decades.

The challenge is that the financial records are split across two very different formats. Records from 1983 to 2022 exist as physical paper documents that were later scanned into PDF files. These include handwritten check registers, typed invoices from forestry companies, journal entry vouchers, and bank statements. More recent records from 2014 to 2025 are stored digitally in QuickBooks. Without a way to bridge these two formats, the county cannot produce a unified financial history.

1.1 Why This Matters

- The county must demonstrate responsible fund management to renew its state timber lease
- Historical spending patterns inform future budgeting and loan decisions
- Manual transcription of 40 years of paper records is impractical without automation
- A reusable pipeline allows the county to process future documents independently

2. Data Sources

The project works with two separate data sources that together cover the full financial history of the timber fund.

	Scanned PDFs	QuickBooks Export
Period	1983 to 2022	2014 to 2025
Format	Scanned PDF images: checks, journal entries, transfer vouchers	Structured CSV/Excel with 13 columns per transaction
Volume	Hundreds of pages across year folders (1983-1990, 1991-2000, 2001-2010, 2011-2014, 2015-2022)	351 transactions
Status	Pipeline built; 1990 document fully tested	Fully cleaned, categorized, and visualized in interactive dashboard

3. Methodology

The project integrates two main pipelines: an OCR-based extraction pipeline for historical scanned documents, and a structured data analysis pipeline for QuickBooks exports. Both pipelines are designed for fully local execution to comply with the county's NDA.

3.1 OCR Pipeline (Historical Records, 1983–2022)

OCR stands for Optical Character Recognition — it is the technology that converts an image of text (like a scanned page) into actual text a computer can read.

The pipeline processes documents through three stages:

- PDF → Images: Each PDF page is converted to a high-resolution image using pdf2image
- OCR: Tesseract reads the image and extracts raw text (imperfect for handwriting)
- AI Extraction: The full raw OCR text is sent to Phi-3 (local AI via Ollama) in a single structured prompt. Phi-3 extracts the date, amount, payee, direction (IN/OUT), category, and confidence level simultaneously, returning a structured JSON array of transactions

3.2 Why Local AI (Ollama/Phi-3) Instead of Cloud AI (Claude/GPT-4)

The NDA with Luce County requires that no financial data leave the county's systems. Cloud AI services like Claude or GPT-4 transmit data to external servers over the internet, which would violate this agreement. Ollama runs the Phi-3 AI model entirely on a local laptop with no internet connection required.

The tradeoff is capability: Phi-3 has 3.8 billion parameters vs. hundreds of billions for cloud models. Phi-3 is good at structured tasks with clear text, but struggles when OCR produces garbled output. This is an honest limitation documented in Section 5.

3.3 Prompt Engineering for Financial Extraction

Rather than using simple keyword rules, the pipeline sends the full page text to Phi-3 with a carefully designed prompt that:

- Identifies the document domain (Luce County timber fund financial records)
- Instructs the model to return structured JSON with specific fields: date, amount, payee, direction (IN/OUT), category, and confidence level
- Provides accounting rules (e.g., Debit = OUT, Credit = IN in journal entries; checks payable TO county = IN)
- Explicitly handles the duplicate problem
- Requests a confidence rating (high/medium/low) so human reviewers know which rows need checking

3.4 QuickBooks Pipeline (Modern Records, 2014–2025)

The QuickBooks export is a structured Excel file with 351 rows, each representing one transaction. The cleaning and categorization process involved:

- Date standardization: Converting text dates to proper datetime format
- Deduplication: Removing 1 exact duplicate row identified
- Missing value handling: 176 missing payees (normal for deposits/transfers), 32 missing accounts, 34 missing memos — all documented
- Four-category classification: Money In, Money Out, Transfer, Other — based primarily on QuickBooks account name, with Type and Payment/Deposit columns resolving edge cases
- Edge case handling: 32 blank-account rows, 11 transfer deposits reclassified as Money In, 8 AP rows with no memos

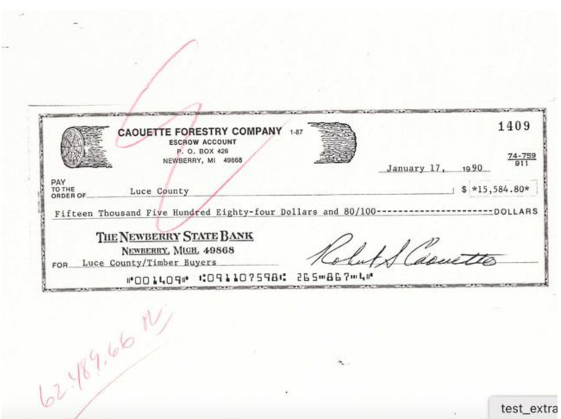
4. Pipeline Testing: 1990 Document (1990_002651.pdf)

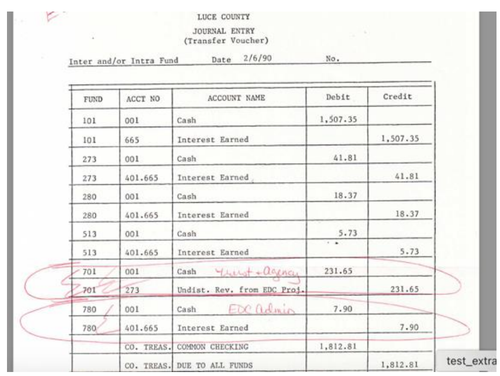
To validate the OCR + AI pipeline and document its realistic capabilities and limitations, the pipeline was tested on a real 1990 Luce County document: 1990_002651.pdf (20 pages, 3.3 MB). This document was chosen because it represents the older, more challenging end of the historical record spectrum.

4.1 Document Types Found Within the PDF

One of the most important findings from this test is that a single PDF file can contain multiple completely different document types, each requiring different extraction logic:

Document Type	Example (Page 1)	Example (Page 2)	OCR Challenge
Physical check	Caouette Forestry Co. check for $15,584.80 to Luce County	—	Cursive signature, handwritten amounts, red pen annotations
Journal entry	—	Transfer voucher with Fund/Acct/Debit/Credit columns, date 2/6/90	Printed table but with handwritten annotations circling specific rows
Payment voucher	—	Page 18: account codes 701 000 / 273 000, amount $29,499	Account codes (e.g., 701 000) misread as dollar amounts


CAOUETTE FORESTRY COMPANY
ESCROW ACCOUNT
P. O. BOX 426
NEWBERRY, MI 49868
1409
January 17, 1990
74-759/911
PAY TO THE ORDER OF Luce County $ *15,584.80*
Fifteen Thousand Five Hundred Eighty-four Dollars and 80/100 DOLLARS
THE NEWBERRY STATE BANK
NEWBERRY, MICH. 49868
FOR Luce County/Timber Buyers
62,489.66


LUCE COUNTY
JOURNAL ENTRY
(Transfer Voucher)
Inter and/or Intra Fund Date 2/6/90 No.

FUND	ACCT NO	ACCOUNT NAME	Debit	Credit
101	001	Cash	1,507.35	
101	665	Interest Earned		1,507.35
273	001	Cash	41.81	
273	401.665	Interest Earned		41.81
280	001	Cash	18.37	
280	401.665	Interest Earned		18.37
513	001	Cash	5.73	
513	401.665	Interest Earned		5.73
701	001	Cash	231.65	
701	273	Undist. Rev. from EDC Proj.		231.65
780	001	Cash EDC Admin	7.90	
780	401.665	Interest Earned		7.90
	CO. TREAS.	COMMON CHECKING	1,812.81	
	CO. TREAS.	DUE TO ALL FUNDS		1,812.81

Fig4.1. Screenshot of Page 1 (check) and Page 2 (journal entry) side by side here

4.2 Extraction Results

Metric	Result	Interpretation
Total pages processed	20	Full document covered
Transactions extracted	41 rows	Multiple transactions per page is normal for journal entries. Note: manual verification (Section 4.3) found the model can under-extract dense tables
Amounts successfully found	39 of 41 (95%)	Only 2 pages returned no data at all
High confidence rows	25 of 41 (61%)	AI flagged its own uncertainty on the remaining 39%. However, manual spot-checking (Section 4.3) found "high confidence" rows can still contain fabricated or incomplete data, so this percentage should not be read as an accuracy rate.
Failed pages	2 (pages 5 and 15)	JSON parse error — AI response was not structured correctly
Date extraction success	Partial — ~60% of pages	Dates in non-standard formats (e.g., 2/6/90) were sometimes missed

4.3 Sample Extraction Output (Page 2 — Journal Entry)

Manual verification against the source document found that Page 2 contains 14 debit/credit line items across 6 funds (101, 273, 280, 513, 701, 780, plus the CO. TREAS. summary lines), but Phi-3 extracted only 4 transactions total. Of those 4, only 1 ($1,812.81) fully matched the source; one amount ($401.67) could not be located anywhere in the original document, suggesting a fabricated value rather than a misread one. A second entry ($231.65) does correspond to a real figure in the source but was mislabeled — the document identifies it as a transfer under "Trust + Agency," not as interest income. All four extracted rows were marked "high confidence," despite the model missing 10 of 14-line items and inventing one figure entirely. This demonstrates that self-reported confidence reflects the model's certainty about what it extracted, not the completeness or accuracy of the extraction. The page's repetitive structure — seven near-identical Fund/Account/Debit/Credit pairs — appears to be a contributing factor: when a document presents many visually similar rows, the model may generate plausible values consistent with the pattern rather than reading each line individually. This finding has a direct practical implication for the county: confidence scores alone should not be used to determine which rows require human review. A more reliable approach is to flag any page with a dense or repetitive table structure for manual spot-checking, regardless of the confidence level Phi-3 assigns.

The table below verifies the extracted output against the source:

Page	Date	Amount	Direction	Category	Confidence	Verified Against Source
2	2/6/90	$231.65	IN	interest_income	High	❌ Amount real (fund 701) but mislabeled
2	2/6/90	$401.67	IN	interest_income	High	❌ Not found anywhere in source — fabricated
2	UNKNOWN	$1,812.81	IN	transfer	High	✅ Confirmed accurate
1	Jan 17, 1990	$15,584.80	OUT*	timber_sale	High	Amount/date correct, direction reversed

5. Results

5.1 QuickBooks Analysis (2014–2025)

The 351 QuickBooks transactions were fully cleaned, categorized, and analyzed. The RLF Payback classification has been confirmed by Tammy Lasley-Henry (Luce County Treasurer) as Money In.

Category	Transactions	Key Accounts	Status
Money In	~146 rows	Interest earned, timber sales	Confirmed
Money Out	~31 rows	Swampland tax, misc expenses	Confirmed
Transfer	~38 rows	Between checking accounts	Confirmed
RLF Payback (AP)	132 rows (38%)	Accounts Payable, Luce County Treasurer	Confirmed Money In — loan repayments back to Township and County

5.2 Key QuickBooks Finding: RLF Payback

132 transactions (38% of all QuickBooks data) sit in the Accounts Payable account but have values in the Deposit column — meaning money is coming in, not going out. The payee for all 132 is "Luce County Treasurer" and memos reference "RLF Payback" (Revolving Loan Fund). Tammy confirmed these are loan repayments flowing back into the timber fund from the Township and County, and they are classified as Money In. An additional 8 A/P deposit rows with no memo remain unidentified and have been flagged.

5.3 Interactive Dashboard

An interactive financial dashboard was built using Reflex, a Python-based web framework. The dashboard runs entirely at localhost:3000 — no data ever leaves the machine. It includes:

- KPI Cards: Total Money In, Total Money Out, Net Total, and Transaction Count
- Yearly Bar Chart: Side-by-side Money In vs. Money Out for each year (2014–2025)
- Monthly Area Chart: Trends over time with seasonal patterns visible
- Account Breakdown: Which QuickBooks accounts contribute most to inflows/outflows
- Year Filter: Dropdown to isolate any single year across all charts

The following image is a screenshot of our Reflex dashboard.

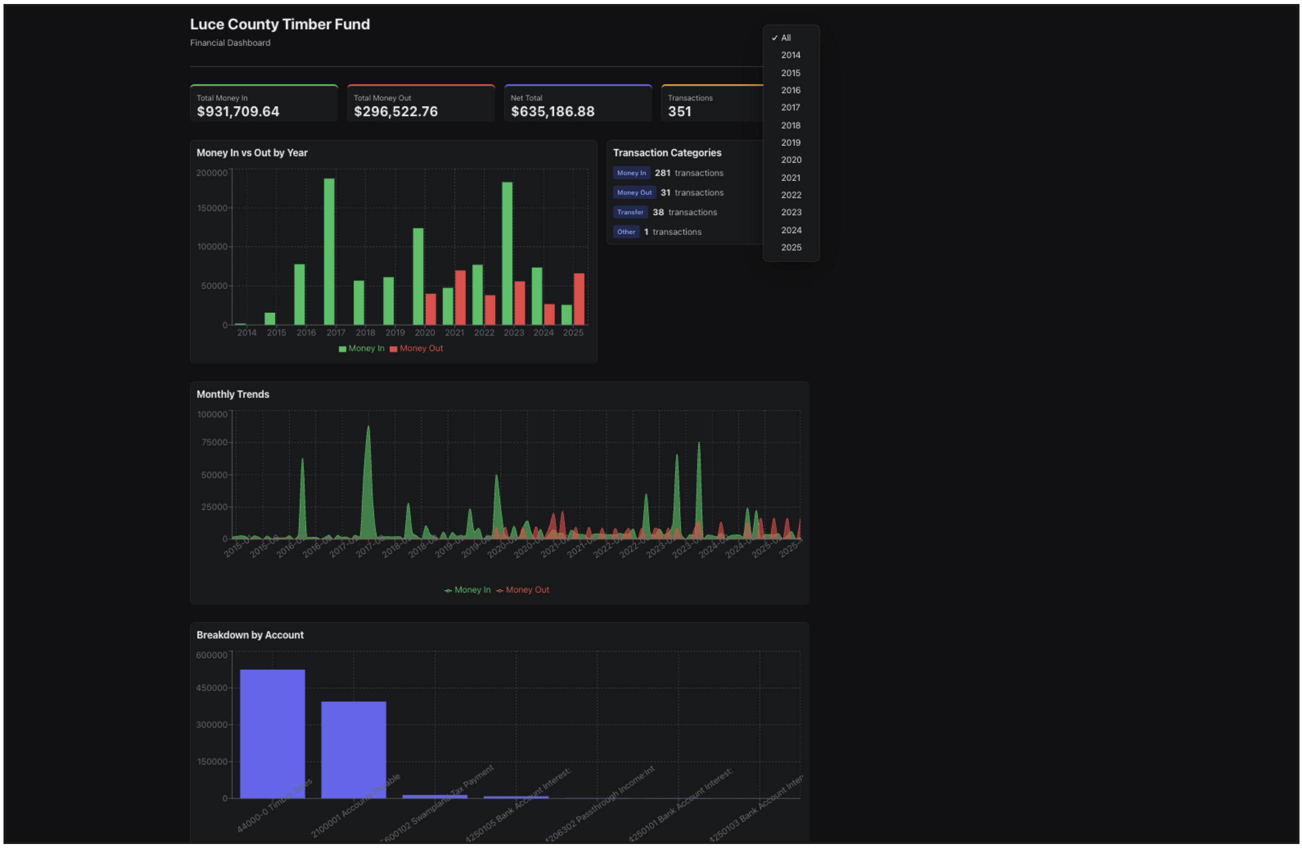


Figure 5.1. Reflex dashboard running locally at localhost:3000, showing real-time KPI cards, yearly Money In vs. Money Out comparison, monthly trend analysis, and account-level breakdown for all 351 QuickBooks transactions (2014–2025). The Year Filter dropdown (top right) allows isolating any single year across all charts.

6. Limitations and Honest Assessment

Being transparent about limitations is essential for the county to use this system safely. The following table summarizes what the pipeline can and cannot reliably do.

Area	What Works	What Doesn't Work Yet
QuickBooks data (2014-2025)	Near-perfect extraction, categorization, and visualization	8 unlabeled A/P deposits needs identification
Printed journal entries	Amounts, dates, directions extracted reliably on simple tables	Account codes sometimes confused with dollar amounts; on dense/repetitive tables, the model can under-extract (missing most rows) and occasionally fabricate plausible-looking values, even at 'high confidence'
Physical checks (scanned)	Large amounts usually detected	Direction (IN vs. OUT) sometimes reversed when both parties are named
Handwritten records (pre-1990)	Some amounts detected if numbers are clear	Cursive text, faded ink, non-standard layouts cause significant OCR failure
Date extraction	Standard formats (MM/DD/YYYY) work well	Short formats like 2/6/90 sometimes missed; dates on cover pages don't apply to all pages

6.1 The Most Difficult Year: 1983–1990

- Document variety: A single PDF from this era can contain physical checks, handwritten journal entries, typed vouchers, and bank registers — each requiring different extraction logic
- Handwriting: Cursive signatures and handwritten amounts produce garbled OCR text (e.g., "Ce Pes a" instead of "Cash")
- Annotations: Red pen notes and circles drawn over printed text interfere with OCR
- No digital reference: Unlike 2015-2025 where QuickBooks provides a cross-check, pre-1990 records have no digital counterpart for validation

6.2 Local AI vs. Cloud AI: The NDA Tradeoff

Cloud-based AI models (Claude, GPT-4) demonstrate significantly higher accuracy on complex financial document interpretation. They can distinguish account codes from dollar amounts, resolve ambiguous direction (IN vs. OUT), and handle garbled OCR text more reliably. However, the NDA requires fully local processing. Phi-3 (3.8B parameters, runs on a laptop) is the correct choice for compliance, with the understanding that it represents a capability tradeoff. Future work could explore larger local models (Llama 3 8B) that offer a middle ground between compliance and accuracy.

7. What to Expect: County Handoff Guide

This section directly addresses what Luce County — specifically Tammy as the county treasurer — should expect when using the deliverables from this project. It is written in plain English, not technical language.

7.1 What You Can Do Right Now

- View 2014–2025 financial trends: Open the Reflex dashboard on a local computer to see yearly Money In vs. Money Out, monthly patterns, and account breakdowns. No technical knowledge required beyond clicking a dropdown.
- Get a structured transaction list: The cleaned_data.xlsx file contains all 351 QuickBooks transactions with categories. Open it in Excel like any spreadsheet.
- Run the pipeline on new documents: When new scanned PDFs arrive, the smart_extraction.py script can process them using the same local AI approach. The output is a CSV file.

7.2 What Still Requires Human Review

- Confidence level alone should not determine which rows get reviewed. Testing found that even rows marked "high confidence" can contain fabricated or incomplete data, particularly on dense or repetitive tables (see Section 4.3). All extracted rows — regardless of confidence level — should be spot-checked against the original document before being used in official reporting, with extra attention paid to pages containing many similar line items (e.g., Account/Debit/Credit rows).
- Amounts from checks (as opposed to journal entries) should be double-checked, as the IN/OUT direction can be misclassified
- Pre-1990 records will need more human review than post-1990 records due to handwriting and document variety
- The 8 unlabeled A/P deposit transactions should be identified with the treasurer before finalizing yearly totals

7.3 What This System Is Not

This pipeline is a first-pass assistant, not an automated auditor. It reduces the time needed to extract and categorize financial records but does not eliminate the need for human judgment. Yearly totals produced by the pipeline should be cross-checked against at least one known year before being used for official reporting or lease renewal documentation.

Think of it this way: instead of manually reading 20 pages from scratch, the treasurer reviews a spreadsheet and corrects ~5–10 flagged rows. That is still a significant time savings, but it is not a fully automated system.

7.4 Sustainability: Running This Independently

The full pipeline can run on any Mac or Windows laptop with:

- Python 3.10+ with the luce Conda environment (all packages documented in requirements.txt)
- Ollama installed with the phi3 model downloaded (free, runs offline)

- Tesseract OCR installed (free, standard installation)

No subscription services, cloud accounts, or internet connection are required for processing. The county owns and controls all the tools.

8. Future Work

The following improvements are recommended for future iterations of this project.

8.1 Image Preprocessing

The single most impactful improvement would be adding image preprocessing before OCR. This includes grayscale conversion, contrast enhancement, noise removal, and page deskewing. For documents with red pen annotations (like many 1990 pages), color filtering can isolate the printed text from handwritten overlays. This would directly improve OCR accuracy and therefore AI classification quality.

8.2 Larger Local AI Model

Replacing Phi-3 (3.8B parameters) with Llama 3 8B or Mistral 7B would provide meaningfully better reasoning on complex financial documents while still running locally. A modern MacBook Pro (M-series) can run these models at acceptable speed.

8.3 Tammy Validation Loop

Building a simple validation interface where the county treasurer can mark extracted rows as correct, incorrect, or uncertain would create labeled training data. Over time, this could be used to fine-tune the local model on Luce County-specific document formats, dramatically improving accuracy for this specific use case.

8.4 Pipeline Remaining Work

- Complete batch processing of 2001–2010, 2011–2014, and 2015–2022 scanned PDF folders
- Reconcile overlapping years between QuickBooks (2015–2025) and scanned PDFs (2015–2022) to avoid double-counting
- Merge both pipelines into a unified yearly summary table covering 1983–2025
- Enhance the dashboard with combined historical + modern data
- Package the tooling with installation documentation for independent county use

9. Acknowledgements

We thank Tammy Lasley-Henry and Shannon Derusha at Luce County for providing access to financial documentation and for their patience and clarity during our working sessions. Their domain knowledge — particularly regarding forestry pricing conventions and the RLF loan structure — was essential for interpreting the financial records correctly.

We thank Dr. Justin Gambrell and Dr. Dirk Colbry at Michigan State University for their guidance throughout this project. We also acknowledge Luce County for providing access to financial documentation under a signed Non-Disclosure Agreement.

Appendix A: Deliverables

Deliverable	Format	Description
smart_extraction.py	.py	OCR + Ollama/Phi-3 extraction pipeline for scanned PDFs. Processes full documents and outputs structured CSV.
QuickBooks Analysis	.ipynb	Step-by-step cleaning, categorization, and summary generation for the 351-transaction QuickBooks export.
cleaned_data.xlsx	.xlsx	351 transactions with Category and Year columns. NDA-protected.
Reflex Dashboard	Python app	Interactive local dashboard at localhost:3000. KPIs, yearly/monthly charts, account breakdown, year filter.
smart_results.csv	.csv	Extraction output from 1990_002651.pdf: 41 transactions with date, amount, direction, category, confidence.

Appendix B: Technology Stack

Tool	Purpose
Python 3.12 / Conda	Core language and environment management (luce environment)
Tesseract OCR	Optical character recognition for scanned handwritten and typed PDFs
Ollama + Phi-3	On-device language model for NDA-compliant transaction classification and extraction
pdf2image	PDF page to high-resolution image conversion for OCR processing
pandas / openpyxl	Data loading, cleaning, grouping, summarization, and Excel I/O
Reflex	Python web framework for the interactive localhost dashboard
MSU GitLab	Private repository: gitlab.msu.edu/xingtia2/luce-lumber (data files excluded per NDA)

Appendix C: NDA Compliance Summary

- All data processing is performed locally on personal machines. No county data is uploaded to public servers or cloud services.
- The interactive dashboard runs on localhost only (port 3000). No external hosting is used.
- The AI classification model (Phi-3 via Ollama) runs entirely on-device with no external API calls.
- Code is stored on MSU GitLab (private repository). Data files (.xlsx, .csv) are excluded via '.gitignore'.

- Notebook outputs are cleared before pushing to GitLab to prevent accidental exposure of financial figures.

Appendix D: Repository Structure

GitLab: gitlab.msu.edu/xingtia2/luce-lumber (private)

Location	Contents
/extractor/	smart_extraction.py, test_extraction.py, extraction_results.csv, smart_results.csv
/luce-dashboard/	Reflex dashboard application (cleaned_data.xlsx excluded)
Root	extraction_test.ipynb (API key must be scrubbed), regular_checking_analysis.ipynb, cleaned_data.xlsx (excluded)
/data/raw/	NDA-approved sample data only. Full records not included.

MICHIGAN STATE UNIVERSITY

UNIVERSITY OUTREACH AND ENGAGEMENT

The MSU Center for Community and Economic Development Regional Economic Innovation (REI) seeks to identify and develop new economic development tools, models, policies, and practices to support innovative economic development, high-growth enterprises, and job creation in distressed regions across the state. REI is establishing a new economic development ecosystem to cope with the ever-changing global and regional dynamics. Through this ecosystem, REI engages innovative and creative minds which results in new economic development practices.

REI was established in 2011. In 2026, REI partnered with the following Michigan State University offices:

Office of the Provost
Office of the VPRI
MSU Jenison Fund
University Outreach & Engagement
College of Communication Arts and Sciences

2026 SLFG

REI

MSU CENTER FOR COMMUNITY AND ECONOMIC DEVELOPMENT
MSU CCED