Can LLMs Credibly Transform the Creation of Panel Data from Diverse Historical Tables?*

Verónica Bäcker-Peral[†] Vitaly Meursault[‡] Christopher Severen[§]

November 2025

Abstract

Multimodal LLMs offer a watershed change for the digitization of historical tables, enabling low-cost processing centered on domain expertise rather than technical skills. We rigorously validate an LLM-based pipeline on a new panel of historical county-level vehicle registrations. This pipeline is estimated to be 100 times less expensive than outsourcing options, reduces critical parsing errors from 40% to 0.3%, and matches human-validated gold standard data with an R^2 of 98.6%. Analyses of growth and persistence in vehicle adoption are statistically indistinguishable whether using LLM or gold standard data. LLM-based digitization unlocks complex historical tables, enabling new economic analyses and broader researcher participation.

Keywords: OCR, Layout Parsing, Entity Linking, Multimodal LLM, Vehicle Adoption

JEL Codes: C80, N72, N32, R40

*We thank Gordon Hanson, Jeffrey Lin, and Allison Shertzer for their helpful comments. Madison Dyhre Hansen, Nassir Holden, Svyatoslav Karnasevych, and Nathan Schor provided valuable research assistance.

[†]Massachusetts Institute of Technology: vbperal@mit.edu

[‡]Federal Reserve Bank of Philadelphia: vitaly.meursault@phil.frb.org

[§]Federal Reserve Bank of Philadelphia: chris.severen@phil.frb.org

1 Introduction

The advent of multimodal Large Language Models (LLMs) catalyzes a watershed moment in extracting historical information for quantitative analysis, particularly by redefining our ability to process complex, heterogeneous historical tables. The tables examined here—early 20th century automobile registration data produced by decentralized state agencies—exemplify heterogeneity through dramatic variation in layout structures, column naming conventions, geographic aggregations, and numeric representation conventions. Such tables lack the standardization common in other contexts (e.g., corporate balance sheets). While table digitization research has increasingly tackled diverse layouts and formats (e.g., Carlson, Bryan, and Dell 2024; Correia and Luck 2023; Silcock et al. 2024; Circi et al. 2024), the salience of heterogeneity in historical administrative sources poses distinct challenges. Not only does successful analysis require accurate cell extraction from highly varied layouts, it also demands extensive cross-document harmonization to reconcile semantic differences and construct cohesive panels. For such tasks, multimodal LLMs offer transformative potential.

We introduce and evaluate a novel digitization architecture to demonstrate the potential of multimodal LLMs. This architecture is specifically suited for transforming historical table scans into cohesive, analysis-ready panels. Unlike approaches that rely on customized machine learning or deep learning models and thus demand significant technical expertise in computer vision or natural language processing, this pipeline leverages the integrated visual and textual understanding of multimodal LLMs to handle complex tabular structures. Crucially, researchers guide the process using natural language prompts, applying their domain expertise regarding the data (e.g., historical reporting conventions, geographical variations) to iteratively refine digitization and harmonization based on observed errors. This domain-knowledge-driven refinement proves particularly advantageous for settings where inconsistencies across source material complicates both extraction and harmonization. Furthermore, the approach has the potential to offer substantial efficiency gains (estimated to be 1/100th the cost of outsourcing with less labor than manual processing), making complex panel dataset creation significantly more accessible to domain experts, regardless of technical background or budget.

To demonstrate the effectiveness of the LLM-driven pipeline, we apply it to a challenging test case: early 20th century county-level vehicle registration tables from the United States. These tables, produced independently by various state-level agencies, exhibit precisely the kind of heterogeneity that makes traditional digitization difficult, thus serving as a useful benchmark. Given the pipeline's sequential complexity (layout, extraction, harmonization) and the risk of compounding errors, rigorous end-to-end evaluation is essential. To enable pipeline development and evaluation, we manually created a "gold standard" dataset from 694 diverse tables (encompassing 49,225 data points). Using separate subsets for prompt development and for overall evaluation (367 and 327 tables), we comprehensively assess performance. Compared to standard OCR solutions often used as a baseline, the LLM pipeline substantially reduces critical layout parsing failures that can render tables unusable (from 40.06% to 0.31%). Furthermore, on the holdout evaluation set, the data

generated by the pipeline match the "gold standard" with high fidelity, achieving an R^2 of 98.6%.

Beyond matching the gold standard at the data-point level, the critical question for applied research is whether any remaining errors propagate and bias downstream statistical inference. We assess this by comparing standard econometric analyses using the LLM data and the gold standard data. The application involves studying local vehicle adoption in the early 20th century, a topic of significant economic and historical importance (Eli, Hausman, and Rhode 2022). Sub-state level analysis has previously been unavailable due to data limitations, making these new insights particularly valuable.[1] We conduct two exercises that reflect common empirical specifications. The first uses lags to study the persistence of vehicle adoption across decades, while the second uses county fixed effects (which could amplify measurement error) to test the relationship between population growth and vehicle adoption. Strikingly, the key regression coefficients estimated using the LLM-based data are statistically indistinguishable from those derived using the manually validated gold standard in both econometric exercises. These analyses reveal interesting patterns in vehicle adoption (very strong persistence and an evolving relationship with population growth over time). Critically, the qualitative and statistical equivalence between results indicates that the LLM pipeline produces data sufficiently robust for complex quantitative analysis and unbiased by the automated digitization process.

This paper contributes to the evolving field of data digitization in five key ways. First, we design, implement, rigorously validate, and evaluate on a holdout sample a complete pipeline architecture leveraging multimodal LLMs specifically to create cohesive panel data from diverse historical table scans. This builds on prior work in both textual extraction and table-specific extraction and extends it into the LLM era (e.g., Shen et al. 2021; Dahl et al. 2023; Stelter and Biehler 2025). Second, we demonstrate how this LLM-based approach makes large-scale table digitization both more cost efficient and significantly more accessible—allowing researchers to leverage domain expertise via prompts rather than technical skill. Third, our pipeline integrates automatic *harmonization*, crucial for handling the heterogeneous table formats common in historical economic sources and often omitted in prior work that focuses on extraction alone.

Fourth, we adapt and formalize the framework of using gold standard data to iteratively improve digitization pipelines to LLMs. Unlike traditional pipelines in which observed errors are translated into opaque technical parameters of OCR and layout parsing systems (like Correia and Luck 2023), domain experts can identify LLM error patterns in natural language and directly adjust prompts accordingly. For example, observing empty cells being filled with hallucinated numbers led to adding the instruction "record all empty cells as empty," while domain-specific knowledge that automobiles counts are often labeled as "passenger cars", "pleasure cars", or "owners" led to adjusting the prompt to acknowledge that "[column] matches are not always textually very similar." This domain-knowledge-driven refinement, conducted through natural language, makes pipeline development accessible to individual researchers, including PhD students, rather than requiring

1. A related innovation, the tractor, has received substantial attention for transforming labor on farms (Olmstead and Rhode 2001; Manuelli and Seshadri 2014), and yet cars were likely even more, and more broadly, transformative (Eli, Hausman, and Rhode 2025).

specialized teams or acquisition of OCR and layout parsing expertise.

Fifth, while gold-standard evaluation is widely used in OCR and extraction studies (e.g., Carlson, Bryan, and Dell 2024; Göbel et al. 2013; Correia and Luck 2023), we extend this paradigm by performing a comprehensive *end-to-end* evaluation proceeding from extraction through harmonization to downstream econometric analysis, explicitly testing equivalence of economic inferences—vital for establishing trustworthiness of AI-processed data given potential error propagation concerns (Battaglia et al. 2024).

Reducing data digitization frictions is crucial because historical data are central to research, allowing scholars to trace the long-run evolution of social phenomena and understand their contemporary implications. From studying how the Dust Bowl shaped agricultural adaptation (Hornbeck 2012) to examining the lasting effects of social connectedness on crime (Stuart and Taylor 2021), or investigating how trade agreements influenced political realignment (Choi et al. 2024), historical data provide crucial insights into current economic and social conditions. Large-scale historical datasets, such as county-level vital statistics (Bailey et al. 2016), demographic records (Haines 2005), and environmental data (Gutmann 2005), enable researchers to analyze long-term patterns and identify relationships that are difficult to establish using only contemporary data (Combes, Gobillon, and Zylberberg 2022). Indeed, reflecting sustained interest, grants related to historical tables from the NSF for Economics alone have totaled approximately $46 million since 2000.[2] Despite such investment, much crucial historical information remains locked in archival documents due to the extensive resources required to convert into machine-readable formats suitable for analysis.

As new tools like the one presented here reduce the cost of accessing this locked-up information, it is crucial that the resulting datasets do not come at the expense of quality or transparency regarding potential errors introduced during automated processing—a concern relevant to both LLM and standard deep learning pipelines. This underscores the importance of adopting rigorous validation and evaluation methods, exemplified by the gold standard approach detailed in this paper, to ensure the suitability and reliability of AI-processed historical data for research.

Extensible to various historical table collections (e.g., decentralized statistical publications, directories, deeds), this LLM-based approach offers an accessible path for researchers across economics subfields due to its low technical barriers and cost-effectiveness. As such, we provide guidance on how researchers can continue to expand the realm of useful data and thus analysis (Abramitzky, Boustan, and Storeygard 2025). By making our code and data publicly available, we aim to facilitate direct use and adaptation, enabling more researchers to unlock information from unstructured documents.[3]

2. Based on NSF Award Search (SES Economics program, keywords "historical," "tables," 2000-present, accessed May 1, 2025).

3. We will make code and data available prior to publication.

Related Literature

The research on automated extraction from tables consists of three broad but related areas: pre-deep-learning pipelines, deep-learning pipelines, and, more recently, prompt-based extraction with multimodal LLMs (Singh and Middleton 2025; Fleischhacker, Kern, and Göderle 2025). Pre–deep-learning models are compared in Göbel et al. 2013. Deep-learning OCR engines and layout models improve accuracy but typically require nontrivial parameter tuning and pipeline engineering (Correia and Luck 2023). More recently, researchers have turned to multimodal LLMs to enable prompt-based extraction across domains, lowering technical barriers and often reducing cost (Balsiger et al. 2024; Circi et al. 2024; Humphries et al. 2025; McLean, Roberts, and Gibbs 2024; Stelter and Biehler 2025).

A close antecedent for our setting, Correia and Luck 2023, emphasizes that out-of-the-box deep learning OCR and layout parsing tools are insufficient and implements an iteratively improved pipeline using human-reviewed ground truth to tune the model parameters. However, they focus on improving digitization accuracy and their setting does not incorporate *cross-document harmonization* (e.g., reconciling geographies, units, and column semantics across heterogeneous sources). In contrast, our approach replaces technical parameter tuning with *prompt tuning* guided by domain expertise and explicitly adds a *harmonization stage*, yielding an *end-to-end* pipeline—from layout recognition and cell extraction to semantic normalization and panel construction—evaluated with a gold-standard and downstream econometric equivalence tests.

Related studies applying deep learning or LLMs to historical documents (Dahl et al. 2023; McLean, Roberts, and Gibbs 2024; Stelter and Biehler 2025; Balsiger et al. 2024; Circi et al. 2024; Humphries et al. 2025) generally emphasize out-of-the-box extraction rather than a formal, domain-informed improvement loop that culminates in harmonized panels. This distinction is central for heterogeneous historical tables, where layout errors and harmonization choices can compound and ultimately affect downstream inference; our evaluation framework in Sections 4.3 and 6 directly addresses this concern.

2 Multimodal LLMs for Historical Table Digitization

Multimodal LLMs differ significantly from traditional OCR methods, which typically handle layout and text extraction in separate, often brittle, stages. Multimodal LLMs integrate vision and language understanding, allowing for a more holistic interpretation of complex document structures, like historical tables with varied layouts and imperfections. This integration also permits guiding digitization using natural language prompts based on domain expertise, unlike the specialized coding needed for traditional tools.

Foundational technologies include the transformer architecture, adapted for both vision and language (Vaswani et al. 2017; Dosovitskiy et al. 2021; Radford et al. 2021). These models typically convert text and images into shared numerical representations (embeddings), enabling reasoning across modalities—connecting, for instance, a table cell's content with its header and visual position.

While architectural specifics vary, their common strength is unified processing of text along with visual document features. See the discussion in Appendix A.1 for more background and detail.

While general-purpose benchmarks for multimodal LLMs are evolving (Liu et al. 2024; Fu et al. 2024), there is an ongoing need for comprehensive benchmarks tailored to specific tasks like digitizing heterogeneous historical tables. Our paper addresses this need by evaluating LLMs on real-world historical economic tables, providing practical insights into their effectiveness at developing panel data for historical economic analysis.

3 Data

The primary dataset comprises publicly available scans of early 20th-century tables from within the U.S. recording annual, county-level vehicle registrations (including vehicle types like cars, trucks, etc.). This source material was produced by various state agencies (e.g., Departments of Transportation or Motor Vehicles), leading to substantial layout and content heterogeneity. This heterogeneity is relatively common with historical data and poses challenges for traditional OCR and layout parsing.

We require high-quality ("gold-standard") data to first develop and validate the prompts we use to interact with the LLMs, and then separately to evaluate overall pipeline performance.[4] The gold standard data consist of 694 tables, which we then randomly split into separate subsets that we use for prompt development and for overall evaluation (367 and 327 tables, respectively). While creating a gold standard dataset is labor intensive, it is necessary for rigorous validation of a data extraction pipeline. Although we created the gold standard dataset upfront, in practice researchers can create this data incrementally during prompt development. Small batches of corrected LLM outputs can inform prompt refinements, creating a virtuous cycle that reduces subsequent manual effort.[5]

We then apply the LLM-based pipeline detailed to the same scanned images as in the gold standard data. This generates an "LLM dataset," which we compare quantitatively to the gold standard dataset to evaluate digitization accuracy. Additionally, we employ the LLM dataset in our empirical analysis, verifying that conclusions drawn remain consistent with those derived using the gold standard data. These tables represent a subset of a larger effort to comprehensively catalog and digitize historical vehicle registration data, potentially encompassing up to 5,000 state-year tables covering the entire 20th century.[6]

4. See Appendix A.3 for additional details about creating and evaluating this dataset.

5. Our methodology in creating the gold-standard data is not particularly novel, and we do not emphasize this as a contribution. Rather, the availability of gold-standard data for prompt refinement and evaluation is key. LLMs offer improvements in the efficiency of creating future gold-standard data.

6. The tables used to create the gold standard consist of the near universe of documents found by our research team as of the end of 2022, when the manual processing began. The dataset continues to grow, but the additional tables are not used in this analysis.

4 Historical Table Digitization Pipeline

We create a multi-stage pipeline that converts scans of diverse historical tables into a cohesive panel dataset, using LLMs for table structure analysis, content extraction, and harmonization. Figure 1a depicts the distinct stages with inputs and outputs at each step. This pipeline is well suited for extending to other datasets but requires iterative prompt development to achieve high performance on new data sources, as we discuss in Section 4.1.

The Image Preprocessing stage takes *Historical Document Images* as input and uses Amazon's Textract API to identify table regions and the Table Transformer (Smock, Pesala, and Abraham 2022) for orientation detection. Tesseract OCR (Smith 2007) is used additionally to confirm table rotation decisions. These established tools require minimal development effort. The output is *Cropped & Oriented Table Images*.

The pipeline then switches to an LLM-based workflow for the next stages. The images first enter the Multimodal LLM Processing stage, where domain knowledge-based prompts guide the LLM in analyzing table structure and extracting content.[7] We develop prompts that first instruct the LLM to act as a careful researcher and avoid hallucinating numbers. Building on this base, we add detailed instructions, for instance, to handle formatting conventions like commas, explicitly record empty cells, and recognize multi-row headers. Header information is carried over for multi-page tables to ensure consistency. State-specific prompts address unique reporting formats; for example, instructing the model that when Illinois data split Cook County into "Chicago" and the remainder, it should label these as distinct entities ("Chicago" and "Cook Excluding Chicago"). We detail and discuss the development of these prompts in Section 4.1. The output of this stage is *Raw CSV Tables*.

The Post-Processing & Alignment stage creates a homogeneous panel dataset from the heterogeneous raw outputs—a common challenge for researchers. A key step is harmonization, which aligns column headers. Here, we again use an LLM, again guided by an iteratively developed prompt, to map extracted field names (e.g., "Cars", "Passenger Cars") to a predefined list of standardized categories (e.g., "Automobiles"). While creating the standardized column list is domain-specific, using an LLM for the mapping is generalizable and accessible. Similarly, structured prompts guide the LLM in standardizing county names against reference lists, using different templates depending on the table format (county-sorted vs. year-sorted). The output is *Aligned Tables* with standardized fields and entities.

The Context-Aware Outlier Detection stage runs automated checks (e.g., population comparisons, time series/cross-column consistency) on aligned (harmonized) data.[8] Unlike the gold standard evaluation, these broadly applicable checks use only readily available data, serving only to flag potential quality issues for researchers without automatic correction.[9] The output is an analysis-ready

7. We use both `claude-3.5-sonnet-20241022` and Gemini `gemini-1.5-pro`; see Section 4.2 for ensembling.

8. In some instances, we also compare the sum of registrations across all counties within a state and to state-level totals published in Federal Highway Administration (Highway Statistics, table MV-201) to check for misalignment.

9. See Appendix A.5 for details.

Cohesive Panel dataset.

Domain-specific knowledge drives processing throughout the pipeline. As shown in the green elements of Figure 1a, we incorporate specialized expertise through carefully crafted LLM prompts (both general and state-specific), structured reference data (column and county lists), and external validation sources (historical population data). These knowledge components emerge from our iterative refinement process: by comparing pipeline outputs against gold standard data, researchers identify key error patterns and articulate the domain knowledge needed to address them. This approach economically focuses human attention on the error patterns that matter most while effectively scaling domain expertise, as the developed knowledge components can then be applied to very large corpora of tables that would be infeasible to process manually.

4.1 Iterative LLM Prompt Improvement

The LLM-based pipeline uses iterative prompt improvement to achieve high accuracy with new datasets (as discussed above and illustrated in Figure 1b). Our approach guides iterative prompt refinement with quantitative error analysis.

The cornerstone of the pipeline's prompt refinement process is validation against the development subset of the gold standard data. Prompt refinement consists of two broad stages.[10] The first stage is exploratory and targets developing a prompt that performs reasonably well across a broad range of settings. As an example, we provide the initial prompt used at the beginning of our exploratory analysis below.

> **Initial Exploratory Prompt**
>
> You are a researcher who carefully digitizes historical statistical tables. You look at scans from old books and put the tables in csv files. Output a csv table. Don't output any other text. The image to process is attached.

Validation in the exploratory stage consists of primarily of digitizing a table in the development data, then visually comparing the table with its gold standard analog. We then alter the prompt to address any observed issues, then re-digitize the table with the updated prompt. If the issue is addressed and no new problems are introduced, we move on to another table. If another problem is introduced, we again alter the prompt to explicitly address the problems.

In our use case, this experimental stage led to the addition of several general purpose statements into the prompt. For example, we added the instruction "record all the empty cells as empty" to the LLM processing prompt after observing that empty cells sometimes contained extraneous symbols like dots. We also found that the guidance "All rows have the same number of columns" helped reduce catastrophic errors (shifted rows and columns).

The second stage of prompt development is a structured workflow that prioritizes improving the prompt to fix observed errors while carefully validating performance so as not to reduce performance

10. The first of these stages need not exist in every pipeline. Our use of the exploratory stage is akin to searching for a region of parameter space the broadly maximizes the likelihoods for maximum likelihood problems, whereas the second stage corresponds to precise optimization.

elsewhere. To do so, we split the development data into 100 random subsamples and use it for development sequentially (i.e., starting with 10 subsamples, increasing to 20, and so on). At each stage, we analyze both aggregate and individual table errors. For aggregate performance, we generate summary statistics (such as those later reported in Table 1) to track overall improvement. At the individual table level, we create difference tables that record discrepancies between the development data and LLM output. This helps identify catastrophic failures, such as shifted rows, versus more isolated errors. We prioritize addressing errors with larger magnitude differences, as these have greater impact on downstream analysis.

This quantitative feedback informs prompt refinement. For example, as we experimented with longer and more detailed prompts that addressed formatting issues, we discovered that the LLM would often return partial tables. We therefore experimented with adding "Output the whole table" to the prompt. After adding this instruction, we re-valiated on the development set to confirm error reduction while ensuring that the new prompt did not inadvertently increase errors elsewhere. This approach capitalizes on researchers' understanding of their sources and context, allowing them to guide the digitization process without requiring specialized technical skills.

The final LLM-Based Workflow involves several calls to the LLM, each with its own goal and prompt, producing distinct intermediate outputs. This modular structure facilitates development and evaluation by allowing us to analyze and debug each component separately. For instance, we can examine raw digitized tables before alignment, assess alignment quality independently, and insert additional processing steps where needed, such as for multi-page tables.

As an example, we provide the final prompt for one of the components within the LLM-Based Workflow. The full set of final prompts is listed in Appendix B.

County- and Year-Sort Prompt (final)

You are a researcher who carefully digitizes historical statistical tables of historical vehicle registration data. You look at scans from old books and put the tables in csv files.

This is a table of historical data from a file called `filename`. The table title is `title`.

Here are some other things to keep in mind:

- Don't make up numbers because those are very important for your research.
- Remove commas from numbers in the tables, the county names and column names.
- Remove dollar signs.
- Don't add decimal points to the numbers.
- Record all the empty cells as empty.
- Output the whole table.
- Empty rows in the image can be represented by multiple dots, encode them as a single blank space, don't add extra columns.
- All rows have the same number of columns.
- The headers can appear in multiple rows and some of the columns might have only some of the header rows. In that case, combine the header rows, starting with the top one, skipping the empty rows for a specific column.
- All columns should have different names and all the rows should have different names.

Output a csv table. Don't output any other text.

The image to process is attached.

4.2 Model Ensembling

After initial development with Claude as the primary LLM, we also processed tables using Gemini with identical prompts. Following established ensembling principles (Dietterich 2000) to improve robustness by mitigating individual model weaknesses, we average results when both models respond, otherwise using the non-missing one. See Appendix A.4 for details on individual model performance.

4.3 Evaluating LLM Performance Using Gold Standard Data

Rigorous evaluation of AI methods requires holdout data. While many researchers traditionally establish rigor by examining model internals, the opacity of many AI models, particularly closed-source LLMs, hinders such scrutiny. Creating holdout gold standard data that reflects the desired output allows for quantitative performance evaluation and aligns with broader scientific machine learning standards (Kapoor et al. 2024).

For historical table digitization, gold standard data can be created by manually correcting outputs from initial digitization attempts (using standard tools, custom software as in our case, or even LLMs with an initial exploratory prompt). While requiring some effort, reliable evaluation can often be achieved with moderately sized gold standard datasets. In our case, performance metrics on the holdout evaluation set stabilize after approximately 150 tables,[11] with additional tables needed for development, making this approach significantly more cost-effective than full-scale manual processing. This gold standard evaluation methodology provides a generalizable framework for ensuring research quality as AI tools become more prevalent in economics.

This evaluation methodology extends beyond table digitization to other AI applications, providing a generalizable framework for ensuring research quality while democratizing dataset creation.

5 Pipeline Performance

We evaluate the performance of the LLM-based digitization pipeline using the holdout (evaluation) subset of the gold standard data. This holdout subset is never used during prompt development (for more details on the sample split, as well as performance comparison on development and holdout sets, see Supplemental Appendix A.6). This approach allows us to quantitatively compare LLMs against traditional layout parsing and OCR tools like Textract and directly answer a crucial question: can LLMs extract historical tabular data at the quality necessary for rigorous economic research? To answer that question we focus on two kinds of errors: structural issues rendering the whole table unusable for downstream analyses, and errors in extracted numbers.

For structural error analysis, we compare the performance of our approach and of Amazon Textract. In our setting, Textract is much more likely to create unusable tables (see Appendix A.2 for examples). We introduce the concept of *critical table parsing errors* that occur when the extraction process fails to produce a structurally sound and analytically useful table. Such errors

11. See Appendix A.6 for convergence analysis.

occur when an extracted table meets any of these conditions: (1) it contains no valid columns that can be used for numerical analysis, where a valid column is one where all cells contain only numeric values; (2) it has "extra cells," meaning some rows contain more content-bearing cells than there are columns defined in the header, indicating structural misalignment; or (3) the table is empty, containing zero rows of data. Failing any condition marks the table extraction as invalid, since such structural issues render the data unreliable for subsequent analysis or require manual intervention to correct before the table could be meaningfully used.

We find substantial differences in critical parsing failure rates. As shown in Table 1a, Textract fails on 40.06% of holdout sample tables, rendering them structurally unusable without manual correction. In stark contrast, the LLM approach shows remarkable robustness, failing on just 0.31% of holdout sample tables.

The LLM-based pipeline also obtains high numerical accuracy. Table 1b reports an impressive R^2 value of 98.6% between true and LLM values, indicating strong fidelity to the original numerical data. While the LLM approach does encounter some challenges, with a total error rate of 21.2% across all extracted data points, it's important to distinguish between two subtypes of numerical errors: missing outputs (8.2%) and incorrect outputs (13.0%). Missing outputs represent a distinct error category not reflected in other accuracy metrics. These errors are less problematic in practice as researchers can fill the values using imputation techniques for panel data manually by looking at the original scans, apply alternative extraction methods, or (continue to) improve LLM output through better prompting.

Among cells where the LLM does produce values, the mean absolute percentage error is just 3.2%, suggesting that when deviations occur, they are typically small. This combination of high structural reliability and numerical accuracy demonstrates that LLMs can extract historical tabular data at a quality level suitable for economic research, especially when compared to traditional layout parsing and OCR-based alternatives. Focusing only on cells with incorrect outputs (Table 1c), the LLM data still achieve an R^2 of 91.5% with true values. Although large outliers exist, most errors are small (median absolute error of 3.0%) with little systematic bias, suggesting suitability for standard outlier treatment.

Since the dataset consists of heterogeneous tables that vary considerably across different states and decades, systematic errors could significantly affect subsequent analyses if performance degraded for particular subsets. Figure 2 examines extraction quality across temporal and geographic dimensions to assess this concern. Results are reassuring: R^2 values remain consistently high across all decades (Figure 2a) and never fall below 97%, with the lowest values observed in the 1950s. Similarly, R^2 values remain above 95% for most states (Figure 2b), dropping below this threshold for only a handful of states. Error rates by decade (Figure 2c) show the highest total error rate, approximately 24% in the 1920s, while the median absolute percentage error for cells with errors remains well-bounded below 8% across all decades. Figure 2d shows more volatility in error rates across states, but importantly, even in states where error rates are higher, the size of the errors (as measured by median absolute percentage error) generally remains below 5%. These patterns

suggest that while performance does vary across contexts, the LLM approach maintains acceptable levels of accuracy across diverse tabular formats and historical periods.

5.1 LLM-Based Pipeline Costs

A significant advantage of the LLM-based digitization pipeline is the potential to substantially lower costs compared to traditional outsourcing solutions. This cost-effectiveness, combined with the simplicity of natural language prompting, positions LLM-based pipelines as a highly attractive alternative, democratizing high-quality digitization for a broad range of economic research applications.

Operational costs are primarily determined by the size of both text and image inputs processed by the LLMs (in our case, Claude 3.5 Sonnet and Gemini 1.5 Pro). To establish a benchmark for comparison, we consider the costs of professional digitization services for similar historical tabular data. Normalizing LLM costs and digitization costs to our sample dataset reveals a striking average cost differential:

- Small-scale outsourcing cost: \$8.24/table.
- Large-scale outsourcing cost: \$6.14/table.
- LLM-based pipeline cost: \$0.03/table.[12]

These figures demonstrate that the LLM-based approach is approximately 100 times less expensive per table than outsourcing alternatives. Additionally, batch processing capabilities can further reduce operational costs (by 50% at the time of writing), widening the cost advantage of LLM-based pipeline. This dramatic reduction makes large-scale digitization financially viable for many research teams.

While initial pipeline setup involves costs for iterative prompt refinement and gold standard creation, the latter can be made more efficient by using LLMs to produce initial gold standard drafts. This approach focuses the manual gold standard creation effort on correcting LLM outputs rather than generating data entirely from scratch, reducing labor costs (see Appendix A.6 for further discussion). These overall setup costs are amortized over the project.

6 Early Automobile Adoption

We evaluate how the processed data perform relative to the gold standard data in two common econometric specifications. We conduct exercises that examine the persistence of automobile adoption between 1920 and 1960 and how it relates to population growth. The first exercise regresses vehicle adoption on lagged vehicle adoption, while the second regresses vehicle adoption on population and includes county-level fixed effects. These two use cases require panel data to estimate serial

12. In our application, about 28% of LLM costs are associated with the length of the input prompt and 72% of the costs are associated with the length of the output.

correlation and historical elasticities in the presence of unit fixed effects. We do not make causal claims about the correlations we report, instead focusing on whether coefficients estimated with the LLM-processed data differ from those estimated with the gold standard data.

Figure 3 plots the resulting data, revealing substantial variation in the levels of vehicle adoption across counties.[13] This figure also reports the population weighted mean of the LLM-based data and compares it to the mean vehicle adoption rate in the U.S. (from Federal Highway Administration, Highway Statistics, table MV-201). The mean LLM-base value closely tracks the national mean. This shows that the pipeline is reasonably accurate in aggregate and suggests that the observed sample of states is broadly representative of the nation.

6.1 Persistence in Vehicle Adoption

We first examine persistence in county-level rates over each decade from 1920 to 1960 by regressing the contemporaneous vehicle adoption rate on the vehicle adoption rate 10 years prior. We denote county-level log vehicle registrations per capita in county c in state s and year t as y_{cst} and estimate:

$$y_{cst} = \rho y_{cs,t-10} + \delta_{st} + e_{cst}, \tag{1}$$

for $t \in \{1930, 1940, 1950, 1960\}$.[14] Specifications also include state-by-year fixed effects, δ_{st}, which play the dual role of isolating within state variation in adoption and controlling for measurement error that may be common to a source document or table or our processing thereof.

In Equation 1, ρ measures county-level persistence (serial correlation) in vehicle adoption. When ρ is close to zero, there is little persistence in county-level vehicle adoption rates between year $t-10$ and year t. When ρ is close to one, persistence is very high across years. This parameter also measures spatial convergence in adoption (e.g., Barro and Sala-i-Martin 1992; Mankiw, Romer, and Weil 1992).[15] For $\rho \in (0, 1)$, growth is slower where initial adoption was higher, indicating convergence towards more similar adoption rates. Adoption diverges if $\rho \geq 1$. Convergence is greater the closer ρ is to 0. Examining ρ's evolution reveals epochs of varying convergence or divergence.

We estimate Equation 1 on two datasets: the LLM-based data and the gold standard data. For these specifications, we restrict the sample to include only those observations that are present in both datasets. The estimate that uses the processed LLM data we label as $\hat{\rho}^{\text{LLM}}$, whereas the estimate that uses the gold standard data we label $\hat{\rho}$. We then stack these models to enable testing whether these coefficients are statistically distinguishable. That is, we adopt a null hypothesis of $H_0 : \rho^{\text{LLM}} = \rho$. To account for within-county error correlation in both datasets, we cluster standard

13. In these exercises, we study "Total Vehicles," which, if not directly reported, we define as the sum of harmonized "Automobiles" and "Trucks" columns, less "Trailers" when present. If the LLM pipeline returns duplicate county-year readings (because of overlap in sources), we systematically select a single value, prioritizing consistency across document vintages and removing infeasible rates (see Appendix A.7 for full details).

14. As Figure 3 shows, the data used for the pipeline and comparison in this paper are very limited prior to 1915, so $y_{cs,1910}$ has no observations.

15. The growth literature often expresses the correlation between contemporaneous and lagged values as a correlation between growth rates and initial values; in our notation, $y_{cst} = \rho y_{cs,t-10} \Leftrightarrow y_{cst} - y_{cs,t-10} = (\rho - 1) y_{cs,t-10}$.

errors at the county level across datasets. A failure to reject the null is evidence that the LLM pipeline produces data that is statistically indistinguishable from the gold standard data in a typical empirical setting.

Panel A of Table 2 indicates substantial persistence in county-level vehicle adoption rates from 1920 to 1960. Estimates of ρ vary from 0.53 to 0.78. Persistence was lowest earlier in the sample. Between 1920 and 1930, $\hat{\rho}^{\text{LLM}} = 0.53$ and $\hat{\rho} = 0.55$, indicating that having adopted 10% more vehicles per capita in 1920 correlates with a bit more than 5% more vehicles in 1930, on average. Conversely, this means that convergence was greatest between 1920 and 1930; growth rates were somewhat slower in areas with greater early adoption. The 1920s were, in aggregate, an era of rapid adoption of vehicles (Norton 2011). The relatively low value of β reflects broad-based increases in ownership that occurred in most US counties.

The following decades exhibited greater persistence in vehicle adoption rates across counties. Estimates of β after 1930 lie between 0.71 and 0.78, indicating that counties with 10% higher adoption in one decade experienced 7%–8% greater adoption in the next decade. Persistence is greatest between 1930 and 1940, during which aggregate vehicle adoption rates flatlined (Figure 3). Romer (1990) observes that the Great Crash reduced registrations, and while production fell sharply in 1930, reduced scrappage rates offset much of this decline (Chow 1957). Reduced investment and greater preservation of existing capital thus likely lay behind the higher local persistence and lower convergence across different counties seen in the 1930s.

The 1940s and 1950s also saw high local persistence in automobile adoption rates, with counties that had greater vehicle adoption maintaining relatively higher adoption levels than other counties, and vice versa. Historically, this period can be divided into two parts: While WWII reduced demand between 1941 and 1945 (Flamm 2006), registrations rebounded post-war, growing rapidly through 1960. Unlike the more rapid convergence in vehicle adoption during the 1920s, aggregate growth in vehicle adoption of the 1940s and 1950s was substantially slower.

This persistence is long lived. Combining persistence coefficients across decades indicates that places with 10% greater vehicle adoption in 1930 (1920) still experienced (4%) 2% greater adoption in 1960. As this time scale is much longer than the typical depreciation schedule of early automobiles, these results suggest that the initial factors that influenced early adoption have had long-lived influence (Brooks and Lutz 2019). The results also suggest that either people continually respond to those early differences (like in Severen and Van Benthem 2022), and/or that factors complementary to vehicles adopted in some places are themselves quite persistent (e.g., Bleakley and Lin 2012; Duranton and Puga 2020).

Crucially for LLM-based pipeline validation, estimates $\hat{\rho}^{\text{LLM}}$ and $\hat{\rho}$ are quantitatively and qualitatively similar across periods (Panel A of Table 2). Tests confirm these estimates are statistically indistinguishable, indicating the LLM-generated data replicate the gold standard patterns sufficiently well for economic and econometric analysis. Furthermore, we find no evidence of systematic bias, as the LLM-based estimates are not consistently above or below the gold standard values.

County-level estimates of the dynamics of vehicle adoption have not been broadly estimable prior to this due to a lack of local panel data.[16] Eli, Hausman, and Rhode (2022) estimate state-level persistence in vehicle registrations from 1919–1929, finding 10 fewer vehicles per 10,000 people in 1919 is correlated with 1.8 percentage point (pp) faster growth in vehicle adoption rates over the next decade. Although they use a different model than Equation 1, adopting their model to our data, we estimate that 10 fewer vehicles per 10,000 people in 1920 is correlated with 0.4–0.5 percentage point faster growth in vehicle adoption rates. Comparing estimates reveals that local convergence is roughly four times slower than state-level convergence. This suggests that the variation in adoption across counties within the same state is much greater than the variation in adoption across states. This underscores the value of our digitization pipeline in developing richer data to broaden economic conclusions.

6.2 Population Growth and Vehicle Adoption

The next empirical exercise examines the correlation between population growth and vehicle adoption and uses panel-unit (county) fixed effects. This type of specification is very common in applied research, and so provides another relevant test of pipeline performance. As before, we do not focus on establishing a causal relationship. Rather, our intent is to test whether economically interesting patterns are equivalent in both the LLM data and our gold standard data.

Specifically, we regress log vehicle registration rate per capita (y_{cst}) on log county population:

$$y_{cst} = \beta \ln(\text{pop}_{cst}) + \alpha_c + \delta_{st} + e_{cst}. \tag{2}$$

This specification includes county fixed effects (α_c) to ensure that β reflects changes in population and vehicle adoption and to limit confounding factors that are time-invariant at the county level, such as locations-specific factors of production (e.g., ports). We estimate this model in ten-year increments to increase the importance of the county-level fixed effects (as they are frequently pivotal in applications); the power of unit fixed effects to control for time invariant factors decreases as panel length increases (Millimet and Bellemare 2023). As before, we include state-year fixed effects.

We limit the sample of counties to those with a population of less than 50,000 in the initial year of each decade in order to isolate the comparison between places that experience rapid growth with places that do not.[17] As in Section 6.1, we estimate the model in Equation 2 on both datasets, restricting the sample to include only observations present in both. Estimates that use the processed data are labeled $\hat{\beta}^{\text{LLM}}$, whereas those from the gold standard data are denoted $\hat{\beta}$. We stack the models in order to test the null hypothesis of $H_0 : \beta^{\text{LLM}} = \beta$, clustering standard errors at the county level across years and models.

The parameter β reflects the (correlational) elasticity of per capita vehicle adoption with population growth (i.e., a 1% population change correlates with a β% change in per capita vehicle

16. One exception is Meir (1981), who studies vehicle adoption in Ohio counties in the 1930s.

17. Note that annual population is linearly interpolated between decades. The same interpolation is used for both calculating per capita vehicle registrations and the independent variable.

adoption). Values of β close to zero indicate that vehicles are being adopted as fast as the population is growing. For $\beta > 0$, vehicle adoption is greater than population growth, whereas people adopt vehicles at a slower rate than population growth for $\beta < 0$. The anticipated sign and magnitude of this correlation depends on how transportation technologies and patterns of land use coevolve in the face of population growth (Cervero and Kockelman 1997; Bento et al. 2005; Ewing and Cervero 2010). If growth is accompanied by investment in non-automobile infrastructure and integrated land use patterns, automobile use may fall. However, if growth is served by roads and land use is segregated, private vehicles may become more necessary. And population growth may itself signify increased economic opportunity and thus income growth, which could increase vehicle adoption (Dargay and Gately 1999).

Panel B of Table 2 shows intriguing variation in the dynamics of the relationship between population growth and vehicle adoption. The coefficient between 1920 and 1930 is near zero, indicating that vehicle registration growth was on par with population growth in the 1920s. The coefficient becomes positive but remains insignificant in the 1930s. Thus, between 1920 and 1940, population growth and vehicle adoption are in sync. This implies that urbanization and local growth were not systematically associated with changes in transportation technology in this period, at least as it pertains to vehicles. This changes after 1940. Between 1940 and 1950, a 10% increase in population is accompanied by a statistically significant 2.7% decline in vehicle adoption per capita. In the following decade, the elasticity is smaller in magnitude but remains significant, indicating a 10% increase in population occurs with a 2.1% decline in vehicle adoption.

Because Equation 2 includes county fixed effects and county land area is fixed, these results can be interpreted as the elasticity between population density and vehicle adoption. Duranton and Turner (2018) provide careful cross-sectional analysis of the relationship between density and vehicle miles traveled in the modern times. They find an elasticity of about -0.07. Our estimates are somewhat different in nature, reflecting historical periods and exploiting changes in population over time within county. Nonetheless, we find estimates both smaller and greater in magnitude than those in Duranton and Turner (2018), reflecting different periods of growth and urbanization.

Crucially, the estimates of $\hat{\beta}^{\text{LLM}}$ and $\hat{\beta}$ in Panel B of Table 2 are statistically indistinguishable in each decade. The p-values of the null hypothesis are all well away from standard statistical significance thresholds. Along with the findings in Section 6.1, this shows that the LLM data are highly accurate in their representation of the gold standard data. This application includes county fixed effects, which should magnify the presence of measurement error. Despite this adversity, the data from the LLM pipeline perform well.

7 Conclusion

Multimodal LLMs offer an effective, accessible, and relatively inexpensive pathway for transforming heterogeneous historical tables into analysis-ready panel data. The LLM-based pipeline architecture we develop and evaluate using historical vehicle registration tables drastically reduces critical parsing

errors compared to standard tools. This approach allows researchers to leverage domain expertise via natural language, lowering technical barriers. The resulting county-level vehicle adoption dataset reveals granular dynamics obscured by state-level data, demonstrating the approach's potential. By dramatically reducing data acquisition costs, such LLM-based methods can fundamentally shift the optimization calculus for data-hungry researchers, enabling research driven more by potential insight than by data accessibility constraints.

As these tools lower the cost of digitizing diverse historical sources, rigorous validation and evaluation become paramount to ensure data reliability. We advocate for accompanying AI-generated datasets with transparent evaluation, using methods like the gold standard approach, to maintain the integrity of research findings. By making our code available, we hope to encourage broader adoption and adaptation of these powerful techniques for unlocking historical information.

References

Abramitzky, Ran, Leah Boustan, and Adam Storeygard. 2025. "New Data and Insights in Regional and Urban Economics." In *Handbook of Regional and Urban Economics,* edited by Dave Donaldson and Stephen J. Redding, 6:715–777. Elsevier.

Bailey, Martha, Karen Clay, Price Fishback, Michael R. Haines, Shawn Kantor, Edson Severnini, and Anna Wentz. 2016. *U.S. County-Level Natality and Mortality Data, 1915-2007: Version 2.*

Balsiger, David, Hans-Rudolf Dimmler, Samuel Egger-Horstmann, and Thomas Hanne. 2024. "Assessing Large Language Models Used for Extracting Table Information from Annual Financial Reports." *Computers* 13 (10): 257.

Barro, Robert J, and Xavier Sala-i-Martin. 1992. "Convergence." *Journal of Political Economy* 100 (2): 223–251.

Battaglia, Laura, Timothy Christensen, Stephen Hansen, and Szymon Sacher. 2024. *Inference for Regression with Variables Generated by AI or Machine Learning.* arXiv: 2402.15585 [econ.EM].

Bento, Antonio M, Maureen L Cropper, Ahmed Mushfiq Mobarak, and Katja Vinha. 2005. "The effects of urban spatial structure on travel demand in the United States." *Review of Economics and Statistics* 87 (3): 466–478.

Bleakley, Hoyt, and Jeffrey Lin. 2012. "Portage and path dependence." *Quarterly Journal of Economics* 127 (2): 587–644.

Brooks, Leah, and Byron Lutz. 2019. "Vestiges of transit: Urban persistence at a microscale." *Review of Economics and Statistics* 101 (3): 385–399.

Carlson, Jacob, Tom Bryan, and Melissa Dell. 2024. "Efficient OCR for Building a Diverse Digital History." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 8105–8115. Bangkok, Thailand: Association for Computational Linguistics.

Cervero, Robert, and Kara Kockelman. 1997. "Travel demand and the 3Ds: Density, diversity, and design." *Transportation Research Part D: Transport and Environment* 2 (3): 199–219.

Choi, Jiwon, Ilyana Kuziemko, Ebonya Washington, and Gavin Wright. 2024. "Local Economic and Political Effects of Trade Deals: Evidence from NAFTA." *American Economic Review* 114, no. 6 (June): 1540–1575.

Chow, Gregory C. 1957. *Demand for Automobiles in the United States.* Amsterdam: North-Holland Publishing Company.

Circi, Defne, Ghazal Khalighinejad, Anlan Chen, Bhuwan Dhingra, and L. Catherine Brinson. 2024. "How Well Do Large Language Models Understand Tables in Materials Science?" *Integrating Materials and Manufacturing Innovation* 13:669–687.

Combes, Pierre-Philippe, Laurent Gobillon, and Yanos Zylberberg. 2022. "Urban economics in a historical perspective: Recovering data with machine learning." *Regional Science and Urban Economics* 94:103711.

Correia, Sergio, and Stephan Luck. 2023. "Digitizing historical balance sheet data: A practitioner's guide." *Explorations in Economic History* 87:101475.

Dahl, Christian M., Torben S. D. Johansen, Emil N. Sørensen, Christian E. Westermann, and Simon F. Wittrock. 2023. "Applications of machine learning in tabular document digitisation." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 56 (1): 34–48.

Dargay, Joyce, and Dermot Gately. 1999. "Income's effect on car and vehicle ownership, worldwide: 1960–2015." *Transportation Research Part A: Policy and Practice* 33 (2): 101–138.

Dietterich, Thomas G. 2000. "Ensemble methods in machine learning." *Multiple Classifier Systems* 1857:1–15.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2021. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* arXiv: 2010.11929 [cs.CV].

Duranton, Gilles, and Diego Puga. 2020. "The economics of urban density." *Journal of Economic Perspectives* 34 (3): 3–26.

Duranton, Gilles, and Matthew A Turner. 2018. "Urban form and driving: Evidence from US cities." *Journal of Urban Economics* 108:170–191.

Eli, Shari, Joshua K Hausman, and Paul W Rhode. 2022. "Transportation revolution: The car in the 1920s." In *AEA Papers and Proceedings,* 112:219–223.

———. 2025. "The Model T." *Journal of Economic History* 85 (1): 110–151.

Ewing, Reid, and Robert Cervero. 2010. "Travel and the built environment: A meta-analysis." *Journal of the American Planning Association* 76 (3): 265–294.

Flamm, Bradley. 2006. "Putting the brakes on 'non-essential' travel: 1940s wartime mobility, prosperity, and the US Office of Defense." *Journal of Transport History* 27 (1): 71–92.

Fleischhacker, David, Roman Kern, and Wolfgang Göderle. 2025. "Enhancing OCR in historical documents with complex layouts through machine learning." *International Journal on Digital Libraries* 26:3.

Fu, Ling, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, et al. 2024. *OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning.* arXiv: 2501.00321 [cs.CV].

Göbel, Max, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. 2013. "ICDAR 2013 Table Competition." In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR),* 1449–1453. Competition Report.

Gutmann, Myron P. 2005. *Great Plains Population and Environment Data: Agricultural Data, 1870-1997 [United States]: Version 1.*

Haines, Michael R. 2005. *Historical, Demographic, Economic, and Social Data: The United States, 1790-2002: Version 3.*

Hornbeck, Richard. 2012. "The Enduring Impact of the American Dust Bowl: Short- and Long-Run Adjustments to Environmental Catastrophe." *American Economic Review* 102, no. 4 (June): 1477–1507.

Humphries, Mark, Lianne C Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray, and Elizabeth Spence. 2025. "Unlocking the archives: Using large language models to transcribe handwritten historical documents." *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* 1–19.

Kapoor, Sayash, Emily M. Cantrell, Katherine Peng, Tan H. Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, et al. 2024. "REFORMS: Consensus-based Recommendations for Machine-learning-based Science." PMID: 38691601; PMCID: PMC11092361, *Science Advances* 10, no. 18 (May): eadk3452.

Liu, Yuliang, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. 2024. "OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models." *Science China Information Sciences* 67, no. 12 (December): 220102. arXiv: 2305.07895 [cs].

Lu, Haoyu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, et al. 2024. *DeepSeek-VL: Towards Real-World Vision-Language Understanding.* arXiv: 2403.05525 [cs.AI].

Mankiw, N Gregory, David Romer, and David N Weil. 1992. "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics* 107 (2): 407–437.

Manuelli, Rodolfo E, and Ananth Seshadri. 2014. "Frictionless technology diffusion: The case of tractors." *American Economic Review* 104 (4): 1368–1391.

McLean, Mark A., David Andrew Roberts, and Martin Gibbs. 2024. "Ghosts and the machine: testing the use of Artificial Intelligence to deliver historical life course biographies from big data." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 57 (3): 146–162.

Meir, Avinoam. 1981. "Innovation diffusion and regional economic development: The spatial diffusion of automobiles in Ohio." *Regional Studies* 15 (2): 111–122.

Millimet, Daniel L, and Marc Bellemare. 2023. *Fixed effects and causal inference.* Technical report. IZA Discussion Papers.

Norton, Peter D. 2011. *Fighting Traffic: The Dawn of the Motor Age in the American City.* Cambridge, MA: MIT Press.

Olmstead, Alan L, and Paul W Rhode. 2001. "Reshaping the landscape: the impact and diffusion of the tractor in American agriculture, 1910–1960." *The Journal of Economic History* 61 (3): 663–698.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. *Learning Transferable Visual Models From Natural Language Supervision.* arXiv: 2103.00020 [cs.CV].

Romer, Christina D. 1990. "The Great Crash and the Onset of the Great Depression." *The Quarterly Journal of Economics* 105 (3): 597–624.

Severen, Christopher, and Arthur A Van Benthem. 2022. "Formative experiences and the price of gasoline." *American Economic Journal: Applied Economics* 14 (2): 256–284.

Shen, Zejiang, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. "LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis." In *Document Analysis and Recognition – ICDAR 2021,* edited by Josep Lladós, Daniel Lopresti, and Seiichi Uchida, 12821:131–146. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing.

Silcock, Emily, Abhishek Arora, Luca D'Amico-Wong, and Melissa Dell. 2024. *Newswire: A Large-Scale Structured Database of a Century of Historical News.* ArXiv:2406.09490, June.

Singh, Loitongbam Gyanendro, and Stuart E. Middleton. 2025. "Tabular context-aware optical character recognition and tabular data reconstruction for historical records." *International Journal on Document Analysis and Recognition* 28:357–376.

Smith, Ray. 2007. "An Overview of the Tesseract OCR Engine." In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition,* 629–633. Washington, DC, USA: IEEE Computer Society.

Smock, Brandon, Rohith Pesala, and Robin Abraham. 2022. "PubTables-1M: Towards Comprehensive Table Extraction From Unstructured Documents." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 4634–4642. June.

Stelter, Robert, and Rafael Biehler. 2025. "Data retrieval from local heritage books—Is artificial intelligence the solution?" Published online 7 Jun 2025, *Historical Methods: A Journal of Quantitative and Interdisciplinary History.*

Stuart, Bryan A., and Evan J. Taylor. 2021. "The Effect of Social Connectedness on Crime: Evidence from the Great Migration." *Review of Economics and Statistics* 103, no. 1 (March): 18–33.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention Is All You Need.* arXiv: 1706.03762 [cs.CL].

Wadekar, Shakti N., Abhishek Chaurasia, Aman Chadha, and Eugenio Culurciello. 2024. *The Evolution of Multimodal Model Architectures.* arXiv: 2405.17927 [cs.AI].

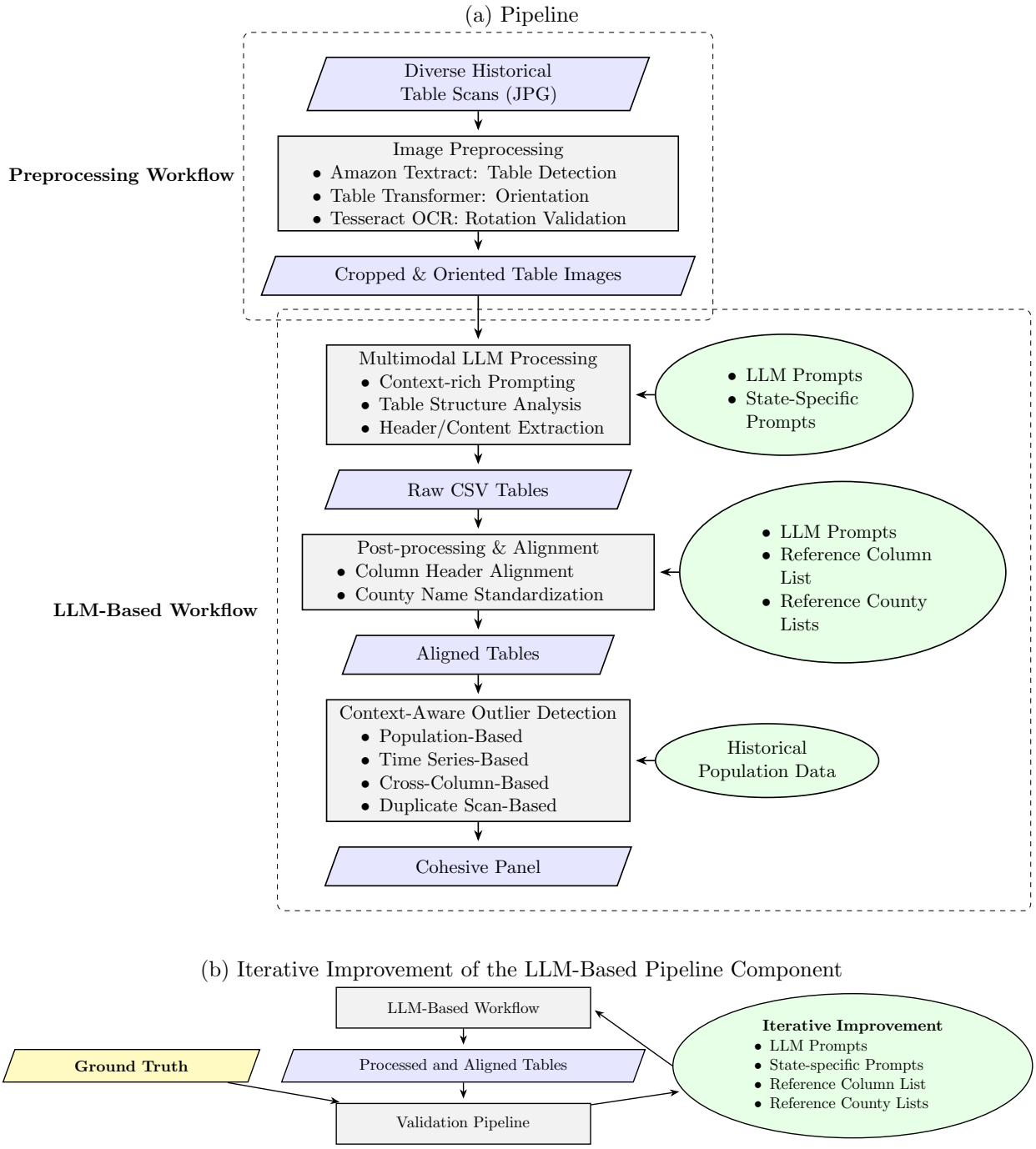


Figure 1: Historical Table Processing Pipeline

This figure illustrates the multi-stage historical table processing pipeline (a) and its iterative improvement cycle (b). In panel (a), the light blue parallelograms represent data inputs/outputs at various stages, grey rectangles depict processing steps or modules, and green ellipses indicate components of domain-specific knowledge (such as LLM prompts, reference lists, and external data) that guide the LLM-based workflow. Panel (b) shows how this LLM-based workflow is refined using ground truth data (yellow parallelogram) to iteratively improve these knowledge components (green ellipse).

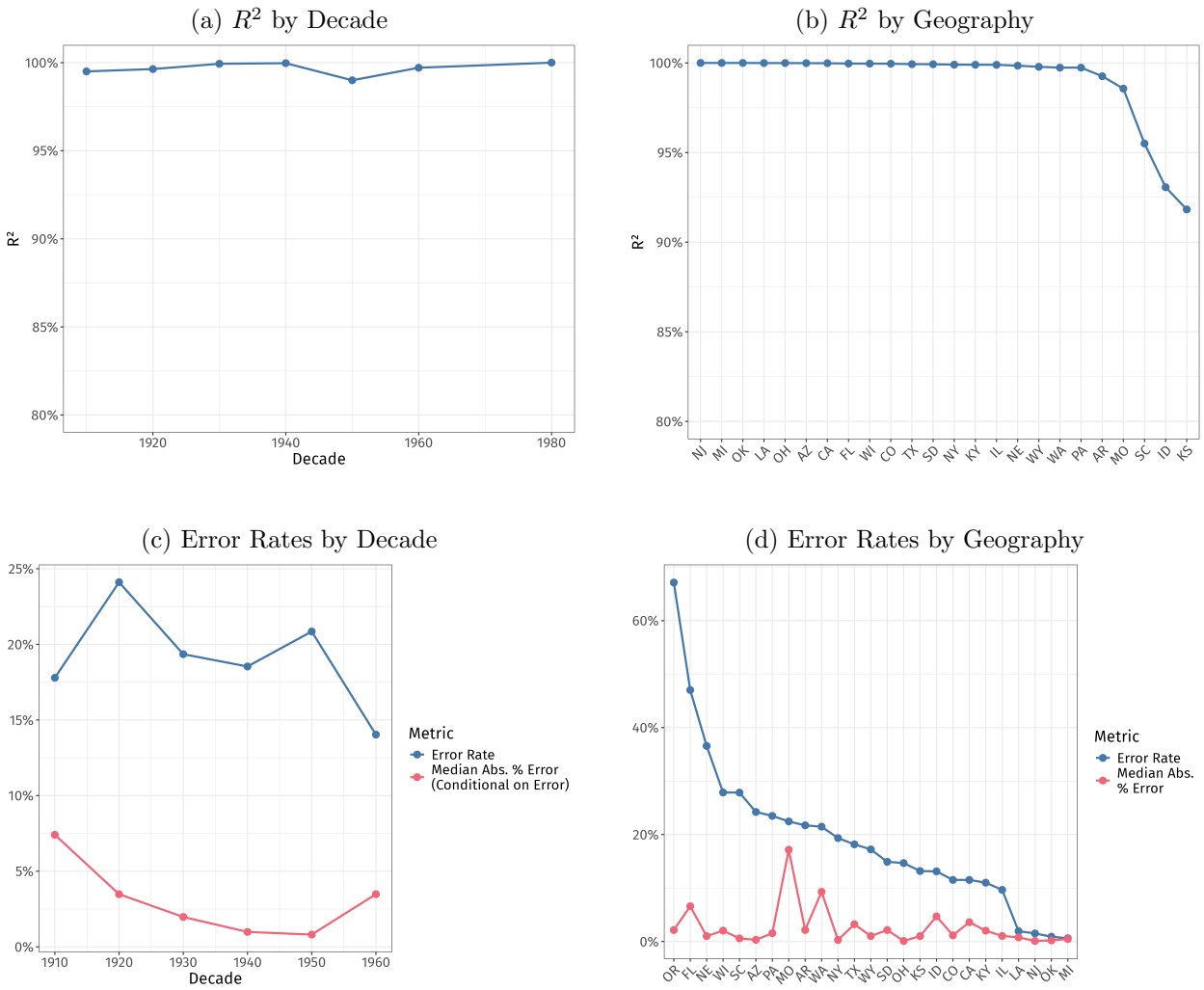


Figure 2: Performance Detail by Time and Geography

These plots examine the LLM pipeline's extraction quality consistency across heterogeneous historical tables from the holdout sample, assessing potential systematic performance variations linked to temporal (decades) and geographic (states) dimensions not evident from aggregate metrics.

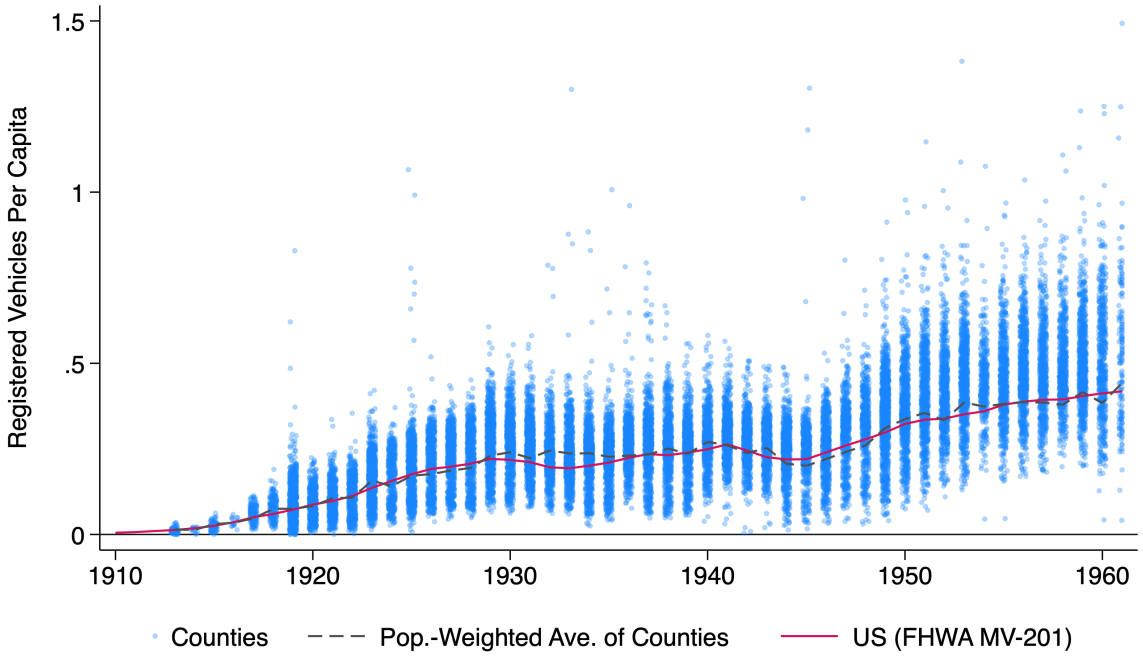


Figure 3: County-Level Vehicle Adoption Rates by Year, 1910–1961 (Cohesive Panel)

This figure plots the distribution of data on vehicles per capita at the county level for each year in our sample. The dashed black line indicates the population-weighted average level of vehicles per capita in our dataset. The solid red line plots the reported national level of vehicles per capita from Federal Highway Administration data.

Table 1: Performance Metrics (Holdout Sample)

(a) Textract and LLM Critical Parsing Failures

Metric	Value
Textract Failure (%)	40.06
LLM Failure (%)	0.31
Num. of Tables	327.00

(b) Overall Performance Metrics

Metric	Value
R^2 (True vs. LLM Values) (%)	98.6
Total Error Rate (%)	21.2
Missing Output (%)	8.2
Incorrect Output (%)	13.0
Mean Error (Units)	-44.0
Mean Abs. Error (Units)	212.8
Mean Error (%)	-0.8
Mean Abs. Error (%)	3.2
Num. of Cells	23067.0
Num. of Tables	327.0

(c) Error Only Performance Metrics

Metric	Value
R^2 (True vs. LLM Values) (%)	91.5
Mean Error (Units)	-311.2
Mean Abs. Error (Units)	1505.2
Median Error (Units)	30.0
Median Abs. Error (Units)	109.5
Mean Error (%)	-5.6
Mean Abs. Error (%)	22.4
Median Error (%)	0.7
Median Abs. Error (%)	3.0
75th Percentile Abs. Error (%)	11.2
95th Percentile Abs. Error (%)	76.2
Num. of Cells	2993.0
Num. of Tables	238.0

This table quantifies LLM pipeline performance on a holdout sample (327 tables, 23,067 cells) not used in prompt development. It compares critical parsing failures against Textract and details overall LLM numerical accuracy, error types, and characteristics of incorrect outputs.

Table 2: Persistence and Correlates of Vehicle Adoption (LLM Data)

	1920–1930		1930–1940		1940–1950		1950–1960	
	y_{cst}^{LLM}	y_{cst}	y_{cst}^{LLM}	y_{cst}	y_{cst}^{LLM}	y_{cst}	y_{cst}^{LLM}	y_{cst}
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Serial Correlation in Adoption								
$y_{cs,t-10}^{\text{LLM}}$	0.530**		0.784**		0.707**		0.763**	
	(0.034)		(0.023)		(0.033)		(0.025)	
$y_{cs,t-10}$		0.554**		0.772**		0.707**		0.773**
		(0.032)		(0.025)		(0.033)		(0.022)
p-value of $H_0 : \hat{\rho}^{\text{LLM}} = \hat{\rho}$	[0.164]		[0.212]		[0.907]		[0.256]	
R^2	0.680	0.704	0.823	0.782	0.752	0.740	0.796	0.802
N	293	293	462	462	336	336	366	366
Panel B. Vehicle Adoption and Population Growth								
$\ln(\text{pop}_{cst})$	-0.008	-0.000	0.081	0.081	-0.270**	-0.283**	-0.212**	-0.199**
	(0.091)	(0.091)	(0.061)	(0.059)	(0.048)	(0.048)	(0.028)	(0.027)
p-value of $H_0 : \hat{\beta}^{\text{LLM}} = \hat{\beta}$	[0.504]		[0.946]		[0.169]		[0.189]	
R^2	0.921	0.943	0.939	0.942	0.938	0.959	0.938	0.961
N	5567	5567	6126	6126	4638	4638	4921	4921

Panel A of this table presents estimates of persistence in vehicle adoption. We estimate a regression of current log vehicles per capita on the 10-year lag of log vehicles per capita in each census year. Panel B presents estimates from a regression of current log vehicles per capita on log population separately for 10-year each period. In both panels, the first column for each period presents results using vehicle data from our LLM-based pipeline, and the second column for each period presents results using our gold-standard dataset. For all time periods, we present p-values for the test of the null hypothesis, $H_0 : \hat{\rho}^{\text{LLM}} = \hat{\rho}$ or $H_0 : \hat{\beta}^{\text{LLM}} = \hat{\beta}$, in square brackets. All specifications include state-year fixed effects and those in Panel B include county fixed effects. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

A Supplemental Appendix

A.1 Multimodal LLM Technical Background

The multimodal LLMs, such as Anthropic's Claude 3.5 Sonnet and Google's Gemini 1.5 Pro used in our pipeline, build upon several key technological components. Text processing relies on transformer-based language models (Vaswani et al. 2017). Input text is converted into numerical tokens via a tokenizer, and these tokens are then mapped to dense vector embeddings that capture semantic meaning.

Visual information is processed by specialized vision models, often adapted from architectures like Vision Transformer (ViT) (Dosovitskiy et al. 2021). These models transform pixel data into numerical representations that encode features ranging from broad structure (like table layouts) to fine details (like cell boundaries and typography). Some models use hybrid encoders processing images at multiple resolutions to capture both aspects effectively.

A critical step is aligning the representations from the text and image modalities. As an example, consider the DeepSeek-VL model (Lu et al. 2024). While not used in our final pipeline, its architecture illustrates common principles. Text is processed by a language model backbone that converts words into numerical tokens via a tokenizer, then transforms these tokens into embeddings—dense vectors capturing semantic meaning where similar words cluster together. Visual inputs are handled by a hybrid encoder system that processes images at different resolutions, transforming pixel data into numerical representations that preserve both broad content (table structure) and fine details (cell boundaries, typography). Text and image representations are then aligned through a two-layer adaptor consisting of multilayer perceptrons—essentially a set of regression-like functions where each input element influences every output element through learned weights with non-linear transformations. This adaptor maps the different visual features into the same dimensional space as the language embeddings, creating a unified numerical framework where both text and image information become compatible vector representations. This mathematical alignment enables the model to effectively process tables by preserving the critical relationships across modalities—connecting the meaning of content with its position in the table grid.

More generally, alignment is often achieved using adapter layers or cross-attention mechanisms. These components project visual features into the same high-dimensional vector space as the text embeddings, creating a unified framework where the model can jointly reason about visual layout and textual content, preserving crucial relationships like the association between a number and its corresponding row and column headers in a table.

Architectural approaches to integrating modalities vary. Wadekar et al. 2024 categorize them into types such as deep fusion (modifying internal LLM layers for cross-modal attention) and early fusion (connecting separate text/vision encoders at the input stage, sometimes after tokenizing images). The specific models used in this paper fall broadly under these integration strategies, enabling the end-to-end processing described in Section 4. For deeper technical surveys and architectural details, see Wadekar et al. 2024 and the documentation for the specific models employed.

A.2 Textract Limitations and LLM Advantages for Historical Table Processing

This appendix provides a concrete example of the key Amazon Textract Layout parsing errors that contribute to its high rates of critical failures discussed in Section 5. Using the 1923 Michigan vehicle registration data shown in Table A1, we demonstrate how structural misinterpretations in OCR processing render extracted data unusable for downstream economic analysis. We then contrast these results with the significantly more accurate output generated by Large Language Models (LLMs) when processing identical source material.

Examining the Textract-extracted data in Table A1b, we observe two key layout parsing failures relative to input table scan in Table A1a. First, Textract incorrectly disjoints rows that should be unified. This is evident in the first data row, where Alcona County's data is split across multiple rows, with the county name separated from its corresponding values. The passenger car count for Alcona (733) appears in a row without a county identifier, while subsequent data shifts position relative to their proper county associations. Second, Textract incorrectly joins values within cells that should be separate. For instance, in the Allegan row, the commercial vehicles count appears as "900 234" instead of the correct value of 909. Similar merging errors occur in the Alpena row, where "2671 1575" appears instead of distinct values in separate cells.

These two error types propagate to create additional structural problems: rows containing numeric data without county identifiers and counties without complete data values. The error in the first row (Alcona) is particularly destructive as it cascades throughout the entire table, causing misalignment between counties and their data points. This single parsing failure shifts all subsequent data values relative to their county identifiers, effectively rendering the entire dataset unusable for economic analysis without substantial manual intervention.

The LLM-structured output in Table A1c demonstrates consistent accuracy in preserving the original table structure. The LLM correctly associates each county with its corresponding data across all columns, maintains proper row boundaries, and accurately distinguishes between adjacent numeric values. This structural coherence avoids the cascading errors observed in the OCR output and produces data suitable for immediate economic analysis.

LLMs are not without transcription errors, however. For example, in Cheboygan County, the LLM incorrectly reads the commercial vehicles value as 178 when it is actually 158—a value correctly captured by Textract. Nevertheless, these isolated numerical errors, which are limited in frequency (as we document in Section 5), are dwarfed by the widespread layout parsing failures associated with Textract that render entire datasets unusable.

While layout parsing and OCR systems like Textract could potentially be optimized to reduce the errors, such optimization requires substantial technical expertise in document processing that is often orthogonal to researchers' core competencies. Each historical table format would likely require different OCR configurations, creating additional workflow complexity when processing diverse archival sources (the core objective of this paper).

In contrast, LLMs demonstrate robust table parsing capabilities with minimal specialized configuration. The high-quality results are achieved largely "out of the box," with additional refinements implemented through prompting techniques that leverage subject matter expertise rather than technical programming knowledge. When data inconsistencies do arise, researchers can address them using their domain knowledge of historical data patterns and structures, working within their established methodological frameworks rather than acquiring tangential technical skills.

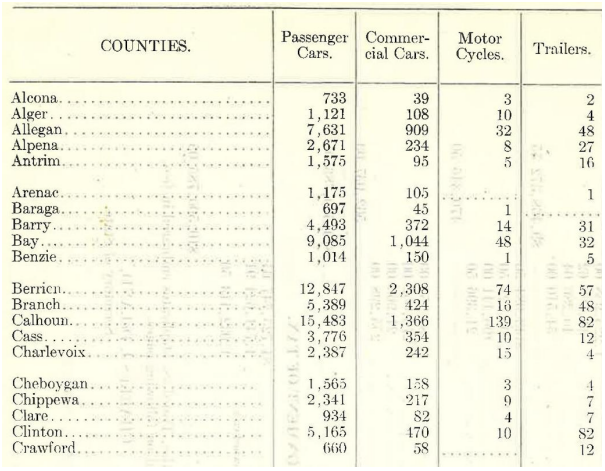
A.3 Creation and Evaluation of the Gold Standard Data

We create the gold standard data using a combination of existing OCR tools (including Textract), custom software, and manual validation. The custom software juxtaposes scanned images with similarly formatted tables and enables rapid manual correction. Multiple passes through this manual digitization step increase accuracy, which we confirm through additional manual evaluation (see below). Gold standard data creation is labor intensive, but we have high confidence in the quality of this dataset.

To provide independent evaluation of the quality of the gold standard data, we randomly selected 100 data points from the gold standard data (10 documents with 10 data points per document) and asked a research assistant to record the values in these cells. Importantly, the research assistant did

Table A1: Michigan Vehicle Registration Data Comparison

(a) Original Michigan 1923 Vehicle Data



COUNTIES.	Passenger Cars.	Commer-cial Cars.	Motor Cycles.	Trailers.
Alcona	733	39	3	2
Alger	1,121	108	10	4
Allegan	7,631	909	32	48
Alpena	2,671	234	8	27
Antrim	1,575	95	5	16
Arenac	1,175	105		1
Baraga	697	45	1	
Barry	4,493	372	14	31
Bay	9,085	1,044	48	32
Benzie	1,014	150	1	5
Berrien	12,847	2,308	74	57
Branch	5,389	424	16	48
Calhoun	15,483	1,366	139	82
Cass	3,776	354	10	12
Charlevoix	2,387	242	15	4
Cheboygan	1,565	158	3	4
Chippewa	2,341	217	9	7
Clare	934	82	4	7
Clinton	5,165	470	10	82
Crawford	660	58		12

(b) Textract-Extracted Vehicle Data

COUNTIES.	Passenger Cars.	Commer- cial Cars.	Motor Cycles.	Trailers.
		39	3	2
Alcona	733	108	10	4
Alger	1121		32	48
Allegan	7631	909 234	8	27
Alpena	2671 1575	95	5	16
Antrim		105		1
Arenac	1175	45	1	
Baraga	697	372	14	31
Barry	4493	1044	48	32
Bay	9085 1014	150	1	5
Benzie				
		2308	74	57
Berricn	12847	424	16	48
Branch	5389	1366	139	82
Calhoun	15483	354	10	12
Cass	3776	242	15	4
Charlevoix	2387			
		158	3	4
Cheboygan	1565	217	9	7
Chippewa	2341	82	4	7
Clare	934	470	10	82
Clinton	5165 660	58		12
Crawford				

(c) LLM-Structured Vehicle Data

COUNTIES	Passenger Cars	Commercial Cars	Motor Cycles	Trailers
Alcona	733	39	3	2
Alger	1121	108	10	4
Allegan	7631	909	32	48
Alpena	2671	234	8	27
Antrim	1575	95	7	16
Arenac	1175	105		1
Baraga	697	59	1	
Barry	4493	372	14	31
Bay	9085	1044	33	32
Benzie	1014	150	1	5
Berrien	12847	2308	74	57
Branch	5389	424	10	48
Calhoun	15483	1366	139	82
Cass	3776	334	10	12
Charlevoix	2587	242	15	4
Cheboygan	1565	178	3	4
Chippewa	2341	217	9	7
Clare	934	82	4	7
Clinton	5165	470	10	82
Crawford	660	58		12

This table visually contrasts the outputs from different digitization methods for the 1923 Michigan vehicle data. It presents the original document scan (a), the structurally flawed extraction by Amazon Textract (b) exhibiting issues like incorrect row splitting and merged cell values, and the corresponding structurally accurate extraction by the LLM (c), which, while not entirely free of numerical transcription errors (e.g., Cheboygan Commercial Cars) preserves overall table integrity.

not observe cell values in the gold standard data. We then compare the cell values input by the RA with those in the gold standard data.

The values precisely matched in 98 out of 100 cells. In the two cells that did not match, it was the gold standard data was correct; the research assistant had incorrectly entered data from an adjacent cell. Thus, careful ex post evaluation reveals that the gold standard data is 100% (in this small random sample). Moreover, this exercise highlights that manual data entry itself is, as always, subject to human fallibility.

A.4 LLM Comparison

In this appendix, we compare the performance of our main LLM, Claude 3.5 Sonnet, with Gemini 1.5 Pro. Before detailing the comparison, it is useful to provide context on our model selection process. We initially tested this pipeline approach with GPT-4 when its image capabilities first became available, but found its performance unsatisfactory for our specific digitization needs. Subsequently, Claude 3 Opus demonstrated excellent performance, achieving accuracy levels very similar to those reported here for Sonnet. However, Claude 3.5 Sonnet achieved this same high performance at a significantly lower cost, leading us to develop and optimize our primary pipeline around this model. We later added Gemini to demonstrate the usefulness of ensembling in the historical table digitization context. It is worth noting that as LLM technology evolves, new models are likely to achieve and exceed these quality levels. Our primary objective in this evaluation is not to identify the single optimal model, but rather to demonstrate that the current generation of LLMs, exemplified by Claude and Gemini, is already sufficient for efficient historical table processing using our methods.

Table A2 mimics the main results presented in Section 5, showing separate metrics for each model. The results demonstrate that Claude outperforms Gemini across key metrics. In terms of overall performance (Table A3), Claude achieves a higher R^2 value of 98.3% compared to Gemini's 97.8%, indicating stronger fidelity to the original numerical data. Crucially, Claude produces far fewer missing outputs (8.2% vs. 18.7%), though Gemini has a lower incorrect output rate (4.7% vs. 11.0%). Overall, Claude's total error rate of 19.2% is superior to Gemini's 23.4%. The mean absolute percentage error for Claude (3.6%) is also substantially lower than Gemini's (14.8%).

When examining only the cells with errors (Table A2b), the performance gap widens dramatically. Claude maintains a high R^2 of 88.8% for error cells, while Gemini's R^2 drops significantly to 65.7%. Claude's mean absolute error in units (1914.5) is substantially lower than Gemini's (5133.3), and Claude's median absolute percentage error (3.9%) is nearly half of Gemini's (7.3%). At the extremes, Claude's 95th percentile absolute error (89.1%) is considerably better than Gemini's (147.5%), indicating Claude produces fewer severe outliers.

The Average column in Table A3 and Table A2b displays the performance of our main specification, demonstrating the benefits of model ensembling through simple averaging of Claude and Gemini outputs. This approach yields an improved R^2 of 98.6% compared to either individual model, suggesting that combining complementary strengths produces more accurate predictions. While the Average model shows a similar missing output rate (8.2%) to Claude, it achieves a better overall error profile with balanced performance across metrics. This exemplifies the principle of ensemble methods as described by (Dietterich 2000), where combining multiple models through weighted voting (in this case simple averaging) can reduce overall error.

It is important to note that these results should be interpreted as representing minimal performance capabilities, as our prompts were specifically developed for Claude. Different models might perform better with specialized prompts tailored to the specific mistakes they make based on our iterative prompt improvement procedure. Nevertheless, these findings clearly establish Claude's superior performance within our experimental framework.

Table A2: Performance Metrics (Holdout Sample)

(a) Overall Performance Metrics

Metric	Model		
	Claude	Gemini	Average
R^2 (True vs. LLM Values) (%)	98.3	97.8	98.6
Total Error Rate (%)	19.2	23.4	21.2
Missing Output (%)	8.2	18.7	8.2
Incorrect Output (%)	11.0	4.7	13.0
Mean Error (Units)	-48.0	-171.8	-44.0
Mean Abs. Error (Units)	228.7	294.0	212.8
Mean Error (%)	-1.1	-13.3	-0.8
Mean Abs. Error (%)	3.6	14.8	3.2
Num. of Tables	327.0	327.0	327.0
Num. of Cells	23067.0	23067.0	23067.0

(b) Error Only Performance Metrics

Metric	Model		
	Claude	Gemini	Average
R^2 (True vs. LLM Values) (%)	88.8	65.7	91.5
Mean Error (Units)	-401.9	-2999.5	-311.2
Median Error (Units)	10.0	195.5	30.0
Mean Abs. Error (Units)	1914.5	5133.3	1505.2
Median Abs. Error (Units)	162.0	300.0	109.5
Mean Error (%)	-9.4	-232.6	-5.6
Median Error (%)	0.3	6.1	0.7
Mean Abs. Error (%)	30.1	258.2	22.4
Median Abs. Error (%)	3.9	7.3	3.0
75th Percentile Abs. Error (%)	21.0	14.7	11.2
95th Percentile Abs. Error (%)	89.1	147.5	76.2
Num. of Tables	228.0	88.0	238.0
Num. of Cells	2528.0	1074.0	2993.0

This table compares the performance of Claude 3.5 Sonnet and Gemini 1.5 Pro on the holdout sample (327 tables, 23,067 cells), using prompts primarily developed for Claude. It details overall metrics (a) and error-only metrics (b) for each model individually and for a simple averaging ensemble ("Average"), highlighting Claude's generally superior performance in this setup and the benefits of ensembling.

A.5 Outlier Detection Based on External Data, Panel Structure, and Column Relationships

To evaluate the quality of data generated by large language models (LLMs), we incorporate outlier detection diagnostics into our pipeline. This approach is valuable not only for assessing the final model quality but also—most crucially—for identifying problematic tables during development that require additional prompting and focused error analysis. The critical difference of these outlier detection methods from gold standard evaluation is that they do not require human-collected validation data and can be applied to tables outside the gold standard dataset. This allows them to serve as valuable quality indicators even when the model is deployed on external data, highlighting potential severe out-of-sample errors that would signal the need for continued model development. Many real-world datasets contain inherent structures and relationships that can be leveraged for such automated validation, providing an efficient mechanism for ongoing quality assurance.

Our outlier detection framework employs four distinct methodologies:

- *Population-based outliers* compare values against county population data from census records. Values are flagged when the ratio of the reported value to county population exceeds a threshold of 2, effectively identifying implausibly large values relative to population size.

- *Time series outliers* examine temporal patterns in panel data structure. We implement a multi-stage detection process that identifies: (1) sharp trend reversals, where significant declines (exceeding 100%) are followed by significant increases or vice versa, for values exceeding 100; (2) anomalous initial observations where changes to subsequent periods exceed 100% and values exceed 500; and (3) anomalous terminal observations with similar characteristics. This approach employs temporal lag structures to compare consecutive observations while filtering out spurious volatility in smaller values.

- *Cross-field outliers* exploit logical relationships between variables within the same table. For

Table A3: Outlier prevalence based on external data (Holdout Sample)

	Model	
Outlier Type	Claude	Gemini
Population Outliers (Count)	21.0	28.0
Population Outliers (%)	0.1	0.1
Timeseries Outliers (Count)	23.0	63.0
Timeseries Outliers (%)	0.1	0.3
Crossfield Outliers (Count)	57.0	88.0
Crossfield Outliers (%)	0.3	0.5
Duplicate Outliers (Count)	223.0	91.0
Duplicate Outliers (%)	1.1	0.5
Total Non-Missing Cells	20396.0	18754.0

This table presents the prevalence of different outlier types (Population, Timeseries, Crossfield, Duplicate) as detected by automated checks on the holdout sample for both Claude and Gemini models. It quantifies these issues in counts and percentages of non-missing cells, offering insights into model-specific error patterns without reliance on human-validated gold standard data.

vehicle registration data, we examine the ratio of automobiles to total vehicle registrations, flagging observations where this ratio is either implausibly low (< 0.3) or logically impossible (> 1.0).

- *Duplicate outliers* leverage repeated measurements of identical data points from multiple sources. When duplicate entries exist for the same county-field-year combination, we calculate their median and standard deviation, flagging points where the ratio of standard deviation to median exceeds 0.5.

- *Duplicate outliers* leverage repeated measurements of identical data points from multiple sources. When duplicate entries exist for the same county-field-year combination, we calculate their median and standard deviation, flagging points where the ratio of standard deviation to median exceeds 0.5.

As shown in Table A3, both models exhibit different outlier patterns. Duplicate outliers represent the most prevalent issue, with Claude showing a higher rate (1.1%) compared to Gemini (0.5%). This suggests Claude produces less consistent results when generating identical values repeatedly. Cross-field outliers appear at comparable rates (Claude: 0.3%, Gemini: 0.5%), indicating both models occasionally generate values violating logical constraints between related fields.

Time series outliers are relatively uncommon (Claude: 0.1%, Gemini: 0.3%), suggesting reasonable temporal consistency in the generated data. Population outliers are the least frequent (0.1% for both models), indicating both models rarely produce values that are implausible relative to population size.

A.6 Determining the Size of the Gold Standard Dataset

Given the crucial role of gold standard datasets in development and rigorous evaluation of LLM-based pipeline, determining how large this gold standard dataset needs to be becomes a central methodological challenge. This section focuses primarily on the holdout portion used for evaluation; the decision of when to stop prompt development is more qualitative, considering both the size

and kinds of errors observed during iterative refinement. While the specific size requirement will inevitably depend on the nature of the digitization task, the practical procedure for evaluating sufficiency remains consistent: researchers can incrementally expand their gold standard dataset and evaluate model performance, observing whether performance metrics converge as dataset size increases.

This iterative approach is illustrated in Figure A1, which tracks various accuracy metrics against the number of tables included. Our gold standard dataset is divided into 100 folds, with 50 folds reserved for developing prompts and the remaining 50 folds designated exclusively for performance evaluation. The convergence analysis below focuses on the 50 evaluation folds. Starting from one fold, we progressively increase the evaluation dataset size by one fold at a time, continuously monitoring changes in performance metrics.

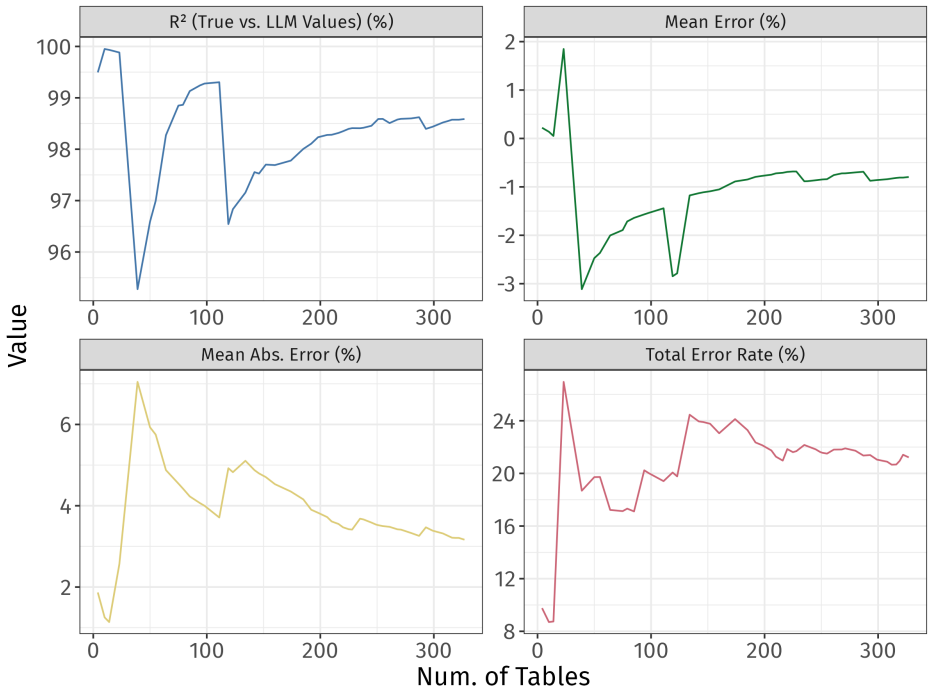
Our analysis shows clear signs of metric convergence as the gold standard dataset expands. Figure A1a demonstrates that the overall R^2 metric between true and LLM-extracted values stabilizes quickly above 98%, even with fewer than 100 tables, suggesting that excellent numerical fidelity is achievable without excessively large gold standard datasets. Metrics related to absolute errors and mean errors similarly demonstrate rapid convergence. For instance, the mean absolute percentage error decreases steadily before stabilizing around 3.2%, indicating diminishing returns from adding further tables. When specifically analyzing cells containing errors (Figure A1b), we observe higher variability initially. The mean absolute error is significantly larger for smaller datasets, reflecting instability in error estimation. However, as we surpass approximately 150 tables, error metrics become markedly stable. The median absolute error, for example, stabilizes around 3.0%.

A key factor influencing the necessary size of a gold standard dataset is the level of heterogeneity present in the data relative to the model's ability to generalize. This consideration is why we examine performance at the state and decade levels explicitly in the main text of the paper (Section 5). If the performance for a specific subsample falls below acceptable thresholds, the gold standard dataset can be strategically expanded with additional data tailored to that subsample, ensuring robust and consistent model performance across diverse conditions.

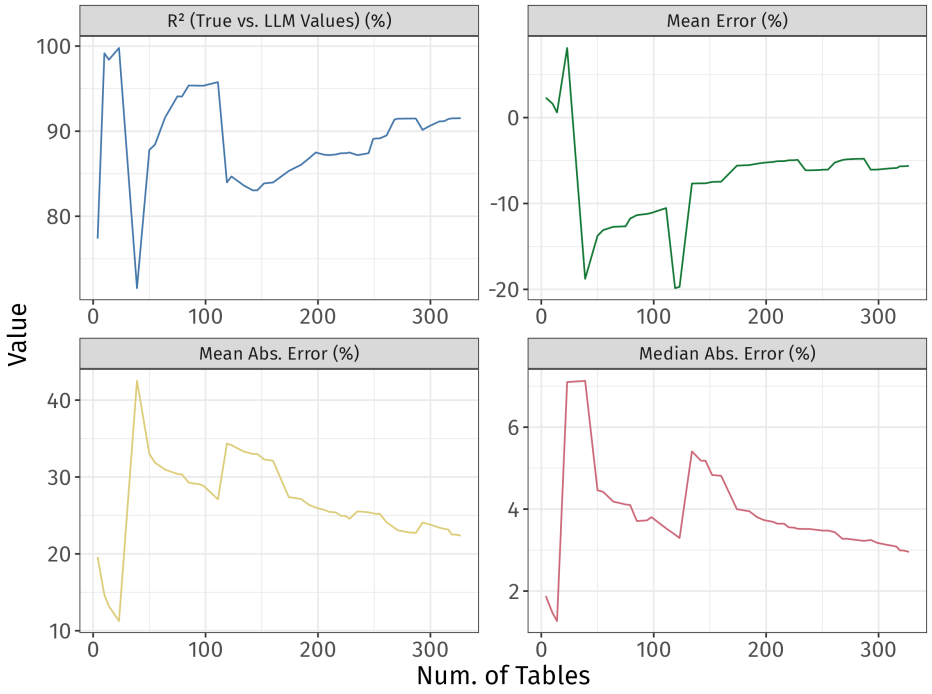
The observed convergence pattern along with the subsample performance statistics support the conclusion that, for this digitization task, a gold standard dataset on the order of a few hundred tables is sufficient to robustly evaluate and establish confidence in the performance of our LLM-based digitization pipeline. While the exact number required may vary across applications, the general practice of incrementally evaluating subsets of data provides a reliable and transparent framework for determining gold standard dataset adequacy in AI-driven economic research.

Although we initially created the gold standard dataset before implementing the full LLM digitization pipeline, our analysis suggests a more cost-efficient methodology. We recommend incorporating gold standard creation into the iterative prompt improvement process discussed in Section 4.1. By initially producing a small number of LLM-processed tables and having researchers correct these outputs, a virtuous cycle develops. As errors are identified and corrected in these initial batches, prompt refinements lead to improved LLM performance on subsequent tables. This approach significantly reduces the manual labor required for later corrections and accelerates the overall digitization timeline while simultaneously building the development portion of the gold standard dataset. Following this procedure, the evaluation portion of the gold standard dataset can be created by manually correcting outputs from the mature LLM-based pipeline, which requires substantially less effort than creating it from scratch at the beginning of the project.

Figure A1: Performance Measurement and the Gold Standard Dataset Size



(a) Overall Performance Metrics



(b) Error Only Performance Metrics

These plots illustrate the convergence of key performance metrics as the size of the gold standard evaluation dataset (number of tables) increases, supporting the determination of sufficient dataset size. Panel (a) shows overall performance metrics like R^2 and mean absolute percentage error stabilizing, while panel (b) tracks error-only metrics, indicating when error characterizations become stable (e.g., median absolute error around 3.0% after $\tilde{1}50$ tables).

A.7 Removing Duplicates

For a substantial share of county-by-year-by-field cells, the data contain multiple readings (a typical example is sequential state publications that print both current and prior year vehicle registrations). We process the duplicates so as to avoid relying on knowledge of the true data. When selecting among duplicates, we prioritize choosing data from documents of similar vintages (sets of tables from a particular state and range of years that share similar characteristics). This empowers state-by-year fixed effects to control for misclassification common to tables from the same source.

We sequentially use the following rules to select among duplicates.

1. Among duplicates, preserve the cell(s) that belongs to more frequent document vintages (groups of documents that are similar in terms of layout and content across years within state).
2. Among duplicates, preserve the cell(s) that can be used to aggregate totals to state-level values.
3. Among duplicates, preserve the cell(s) that returned values from a greater number LLM models.
4. For duplicates which return values from multiple LLMs, preserve the cell(s) for which the values across LLMs match.
5. For duplicates that can be aggregated to a state-level total, preserve the cell(s) that contributes to the most accurate state-level total.
6. Among duplicates, preserve the cell(s) that is closest to the state-level per capita vehicle adoption rate.
7. Among duplicates, preserve the cell(s) that has a corresponding gold-standard value.
8. Arbitrarily choose the between remaining duplicates (using the cell with the lowest document ingestion number).

B LLM Prompt Collection

Below we present the the full, final collection of prompts used to generate digitized and harmonized tables. We settled on the following ordered structure for LLM calls:

1. **Multiple Tables** identifies whether multiple tables in the same scan are part of one big table, or whether they should be processed separately.
2. **County- and Year-Sort Prompt** is the main prompt that converts the JPG into a CSV table.
    - **Prompt for Parts** is added to this prompt if this page is part of a bigger table.
    - **County-Sort Column Header Instructions** and **Year-Sort Column Header Instructions** provide additional instructions that are appended onto this prompt.
3. **Prompt Columns** and **Prompt Counties** are then used to align the column and county names to a provided set of potential column and county names.

Multiple Tables Prompt

> The image you see is a scanned record containing historical data on vehicle registrations. It contains `num_tables` tables, identified in different colors and labelled as `table_list`. You want to identify if any of the tables can be stacked on top of each other. Tables that can be stacked have the following traits:
>
> - They have the same number of columns.
> - Their columns refer to the same thing.
>
> Please output a list containing the numbers of the tables which can be stacked, in the order that they should be stacked. If none of the tables should be combined, output a blank list. Please also output the list first, with nothing before it. Afterwards, output your reasoning.

County- and Year-Sort Prompt

> You are a researcher who carefully digitizes historical statistical tables of historical vehicle registration data. You look at scans from old books and put the tables in csv files.
>
> This is a table of historical data from a file called `filename`. The table title is `title`.
>
> Here are some other things to keep in mind:
>
> – Don't make up numbers because those are very important for your research.
> – Remove commas from numbers in the tables, the county names and column names.
> – Remove dollar signs.
> – Don't add decimal points to the numbers.
> – Record all the empty cells as empty.
> – Output the whole table.
> – Empty rows in the image can be represented by multiple dots, encode them as a single blank space, don't add extra columns.
> – All rows have the same number of columns.

- The headers can appear in multiple rows and some of the columns might have only some of the header rows. In that case, combine the header rows, starting with the top one, skipping the empty rows for a specific column.
- All columns should have different names and all the rows should have different names.

Output a csv table. Don't output any other text.

The image to process is attached.

Prompt for Parts

The table you see is part of a bigger table. If the column names for the table are not provided, use the following column names in this order: `headers`.

County-Sort Column Header Instructions

Please format the column headers that do not refer to a geographic levels (such as counties) as `Field: ENTER_FIELD; Type: ENTER_TYPE; Date: ENTER_DATE`. The field should by the type of data that the column refers to (e.g. automobiles, trucks, population). The type should either be `Vehicle` if the column refers to a number of vehicles, `Fee` if the column refers to a dollar amount, and `Other` if the column refers to another amount. The date is the date to which the column refers to, and should be formatted in one of three ways: (1) if the column refers to a full calendar year, enter only the year formatted as `YYYY`, (2) if the column refers to a period between two dates, format it as `MM/DD/YYYY to MM/DD/YYYY`, and (3) if the column refers to a number of months within a year, format it as `YYYY (XX months)`, where `YYYY` is the year, and `XX` is the number of months in the year that the data refers to.

Use the column headers in the table and the table title to determine the field, type and date for each column. If the table has multiple headers for each column, make sure you merge them. If one of the columns refers to a geographic level, such as US counties or cities, relabel the column header as `Geography`.

Year-Sort Column Header Instructions

Please format the column headers that do not refer to the date as

`Field: ENTER_FIELD; Type: ENTER_TYPE; Geography: ENTER_GEOGRAPHY`. The field should by the type of data that the column refers to (e.g. automobiles, trucks, population). The type should either be `Vehicle` if the column refers to a number of vehicles, `Fee` if the column refers to a dollar amount, and `Other` if the column refers to another amount. The geography is the geographical area to which the column refers to, which could be a county, city or state. Put the full name of the geographic area.

Use the column headers in the table and the table title to determine the field, type and geography for each column. If the table has multiple headers for each column, make sure you merge them. If the first column refers to US counties, label it `Counties`. If one of the columns refers to a date, the column should be titled `Date` and it should be formatted in one of three ways: (1) if the row refers to a full calendar year, enter only the year formatted as `YYYY`, (2) if the row refers to a period between two dates, format it as `MM/DD/YYYY to MM/DD/YYYY`, and (3) if the row refers to a number of months within a year, format it as `YYYY (XX months)`, where `YYYY` is the year, and `XX` is the number of months in the year that the data refers to.

Columns Prompt

You are a researcher aligning the column names of OCR'd tables with column names processed by humans. You will be given two lists of column names: Assess and True. Please output a python dictionary where every column in Assess is assigned a corresponding column in True. Don't output any other text.

Here are some things you should know:

There might be columns in Assess that are not in True and there might be columns in True that are not in Assess.

In that case, assign `NoCorrespondingColumnX` to that column, where `X` is the number of column with that label.

The matches are not always textually very similar. For example, Autombiles are often recorded as `Cars`, `Passenger Cars`, `Owners`, `Pleasure Cars` etc.

Assess: `{cols_yhat}`

True: `{cols_y}`

Counties Prompt

You are a researcher aligning the county names of OCR'd tables with county names processed by humans. You will be given two lists of counties names: Assess and True. Please output a python dictionary where every county in Assess is assigned a corresponding county in True. Don't output any other text.

Here are some things you should know:

- Every county in True can appear in Assess only once.
- Assign None to the county in Assess if it's not in True.

Assess: `{str_counties_yhat}`

True: `{str_counties_y}`