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

Verónica Bäcker-Peral† Vitaly Meursault‡ Christopher Severen§

May 14, 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 100 times less expensive than outsourcing, 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: C40, C80, N72, N32, R40

*We thank Gordon Hanson, Jeffrey Lin, and Allison Shertzer for their helpful comments.

Disclaimer: This paper represents research that is being circulated for discussion purposes. The views expressed in this paper are solely those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. All errors or omissions are the responsibility of the authors.

†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 economic analysis, particularly by redefining our ability to process complex, heterogeneous historical tables. These tables—which typically feature non-standard layouts and formats and are often produced by varied, decentralized sources—embed rich quantitative data but pose distinct challenges compared to the large volumes of textual data tackled by earlier deep learning tools (e.g., Shen et al. 2021; Carlson, Bryan, and Dell 2024; Arora et al. 2024; Silcock et al. 2024). Successfully leveraging such tables requires overcoming hurdles in both accurate digitization and effective harmonization to assemble the cohesive panel datasets needed for rigorous analysis. For such tasks, 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 heterogeneous historical table scans into cohesive, analysis-ready panels. Unlike approaches that rely on specialized deep learning models for each step and thus demand significant technical expertise in computer vision or natural language processing (especially when facing varied layouts), this pipeline leverages a multimodal LLM's integrated visual and textual understanding to handle complex and varied layouts. 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 iterative, domain-knowledge-driven refinement proves particularly advantageous for the heterogeneous table structures pervasive in historical economic sources. Furthermore, the approach offers substantial efficiency gains (in our case, 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 and validate the effectiveness of the LLM-driven pipeline, we apply it to a challenging test case: early 20th century U.S. county-level vehicle registration tables. These tables, produced independently by various state-level agencies, exhibit precisely the kind of heterogeneity that makes traditional digitization difficult, thus serving as an ideal benchmark. Given the pipeline's sequential complexity (layout, extraction, harmonization) and the risk of compounding errors, rigorous end-to-end validation is essential. To enable such robust quantitative evaluation, we manually created a "gold standard" dataset from 694 diverse tables (encompassing 49,225 data points). Using separate subsets for development and 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 matches the "gold standard" with high fidelity, achieving an R^2 of 98.6%.

Beyond matching the gold standard at the data-point level, a crucial test is whether the LLM-generated data yield reliable results in downstream applications. 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 importance (Eli, Hausman, and Rhode 2022) where accessing sub-state data via tables like these offers unique insights.[1] We conduct two exercises that reflect common empirical specifications: one using lags to study the persistence of vehicle adoption across decades, and another using 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. While the analyses themselves reveal interesting patterns in vehicle adoption (such as strong persistence and a changing relationship with population growth over time), the critical finding for our purposes is this qualitative and statistical equivalence. It demonstrates that the LLM pipeline produces data sufficiently robust for complex quantitative analysis, yielding economic inferences that are not significantly biased by the automated digitization process.

This paper contributes to the burgeoning field of data digitization in four key ways. First, we design, implement, and rigorously validate a complete pipeline architecture leveraging multimodal LLMs specifically to create cohesive panel data from diverse historical table scans, addressing challenges distinct from primarily textual extraction (e.g., Shen et al. 2021; Carlson, Bryan, and Dell 2024). Second, we demonstrate how this LLM-based approach makes large-scale table digitization significantly more accessible—allowing researchers to leverage domain expertise via prompts rather than technical skill—and dramatically more cost-effective (approximately 1/100th the cost of outsourcing). Third, our pipeline integrates automatic harmonization, crucial for handling the heterogeneous table formats common in historical economic data sources. Fourth, building on and extending recent gold-standard validation approaches used for OCR (e.g., Carlson, Bryan, and Dell 2024), we perform a comprehensive end-to-end pipeline validation using a manually created gold standard, explicitly assessing the impact on downstream econometric analysis. Such rigorous evaluation is vital for establishing the trustworthiness of AI-generated historical data in economics, particularly given potential pitfalls where complex pipelines might introduce errors affecting inference (Battaglia et al. 2024), a concern our validation framework directly addresses.

Reducing data digitization frictions is crucial because historical data are central to economic research, allowing scholars to trace the long-run evolution of economic 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. Indeed, reflecting sustained interest, grants related to historical tables from the NSF for Economics

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).

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 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 in economics and related fields 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,[3] we aim to facilitate direct use and adaptation, enabling more researchers to unlock information from unstructured documents.

2 Multimodal LLMs for Historical Table Digitization

Multimodal LLMs differ significantly from traditional Optical Character Recognition (OCR) methods, which typically handle layout and text extraction in separate, often brittle, stages. Multimodal LLM 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 (see Appendix A.1; Wadekar et al. 2024), their common strength is unified processing of text along with visual document features.

Standardized benchmarks are evolving (Liu et al. 2024; Fu et al. 2024), but performance on complex tasks like digitizing heterogeneous historical tables has not been comprehensively studied. Our paper addresses this gap by evaluating LLMs on real-world historical economic tables, providing practical insights into their effectiveness at developing panel data for historical economic analysis.

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.

3 Data

The primary dataset comprises scans of early 20th-century tables from within the U.S. recording annual, county-level vehicle registrations (including vehicle types like cars, trucks, etc.). Produced by various state agencies (e.g., Departments of Transportation or Motor Vehicles), these tables exhibit considerable layout and content heterogeneity, posing challenges for traditional OCR and layout parsing.

For rigorous evaluation, we created a "gold standard" dataset (694 tables, 49,225 data points) by manually correcting numeric and formatting errors in initial AWS Textract outputs using purpose-built software. This dataset forms the basis for the quantitative performance assessment described in Section 5.

We then apply the LLM-based pipeline detailed in this paper to the same scanned images. 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 694 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. 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.[4] The gold standard validation demonstrates the effectiveness and suitability of the LLM-based pipeline for this extensive digitization project.

4 Historical Table Digitization Pipeline

We combine LLMs with traditional computer vision techniques to create a multi-stage pipeline that converts scans of diverse historical tables into a cohesive panel dataset. 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 improvement 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 (Smith 2007) OCR is used additionally to confirm table rotation decisions. These established tools required minimal development effort. The output is *Cropped & Oriented Table Images*.

These images then enter the Multimodal LLM Processing stage, where domain knowledge-based prompts guide the LLM in analyzing table structure and extracting content.[5] We developed detailed prompts instructing the LLM, for instance, to act as a careful researcher, avoid hallucinating numbers, handle formatting conventions like commas, explicitly record empty cells, and recognize

4. The dataset continues to grow, but the additional tables are not used in this analysis.

5. We use `claude-3.5-sonnet-20241022` and Gemini `gemini-1.5-pro`. See Section 4.2.

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 splits Cook County into "Chicago" and the remainder, it should label these as distinct entities ("Chicago" and "Cook Excluding Chicago"). 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 economists. A key step is harmonization, which aligns column headers. Here, we again use an LLM, guided by a structured 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 component runs automated checks (e.g., population comparisons, time series/cross-column consistency) on aligned (harmonized) data.[6] Unlike gold standard evaluation, these broadly applicable checks use no human-validated data, serving only to flag potential quality issues for researchers without automatic correction. See Appendix A.4 for details.

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 serve as the primary targets for our iterative refinement process, allowing us to systematically encode domain expertise into the automated workflow.

4.1 Iterative LLM Prompt Improvement

As illustrated in Figure 1b, LLM-based pipelines require iterative prompt improvement to achieve high accuracy with new datasets. Our approach focuses on iterative prompt refinement guided by quantitative error analysis.

The cornerstone of our improvement process is a gold standard dataset. We split this dataset into 100 random subsamples and process it cumulatively—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. For individual tables, we create difference tables that record discrepancies between gold standard data and LLM output. This helps identify catastrophic failures such as shifted rows versus more isolated errors. We prioritize addressing errors with larger magnitude

6. 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.

differences, as these have greater impact on downstream analysis.

This quantitative feedback directly informs prompt refinement. For example, the instruction "record all the empty cells as empty" in the LLM processing prompt was added after we observed that empty cells sometimes contained hallucinated numbers, shifted content, or extraneous symbols like dots. After adding this instruction, we re-evaluated on the gold standard set to confirm error reduction while ensuring that the new prompt did no inadvertently increase errors elsewhere. This approach capitalizes on economists' and domain experts' understanding of their sources and context, allowing them to guide the digitization process without requiring specialized technical skills.

This iterative process differs significantly from standard deep learning refinement, which requires modifying code using specialized technical knowledge. In contrast, LLM-based workflows rely on natural language prompting, leveraging economists' existing expertise in interpreting historical tables.

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.3 for details on individual model performance.

4.3 Gold Standard Evaluation: A Rigorous and Generalizable Framework for AI in Economic Research

Rigorous evaluation of AI methods for economic measurement often necessitates a gold standard dataset approach. While economists traditionally establish rigor by examining model internals, the opacity of many AI models, particularly closed-source LLMs, hinders such scrutiny. Creating a gold standard dataset reflecting the desired output allows for quantitative performance evaluation independent of model interpretability, aligning with broader scientific machine learning standards (Kapoor et al. 2024).

For historical table digitization, such datasets can be created by manually correcting outputs from initial digitization attempts (using standard tools or LLMs). While requiring some effort, reliable evaluation can often be achieved with moderately sized datasets, making this approach significantly more cost-effective than full-scale manual processing or outsourcing. In our data, performance stabilizes at after 150 tables.[7] 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 in economics, providing a generalizable framework for ensuring research quality while democratizing dataset creation.

7. See Appendix A.5 for convergence analysis.

5 Pipeline Performance

We evaluate the performance of the LLM-based digitization pipeline using a holdout subset of the gold standard data. This holdout subset was not 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.5). 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, 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 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%.

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 types of 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 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), only dropping below this threshold for 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 its substantially (up to 100x) lower cost 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.[8]

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

8. 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.

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.5 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 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.[9] 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 $\tau \in \{1930, 1940, 1950, 1960\}$.[10] Specifications also include state fixed effects, $\delta_{s\tau}$, which play the dual role of isolating within state variation in adoption and controlling for measurement error that may 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,

9. 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.6 for full details).

10. 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.

and Weil 1992).[11] 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. We label the estimate that uses the processed LLM data 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 $H_0 : \hat{\rho}^{\text{LLM}} = \hat{\rho}$. To account for within-county error correlation in both datasets, we cluster standard 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,

11. The growth literature often expresses the correlation between contemporaneous and lagged values as a correlation between growth rates and initial values: our model $y_{cs\tau} = \rho y_{cs,\tau-10}$ is equivalent to $y_{cs\tau} - y_{cs,\tau-10} = (\rho - 1) y_{cs,\tau-10})$.

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.[12] Eli, Hausman, and Rhode (2022) estimate state-level persistence in vehicle registrations from 1919–1929, finding 10 fewers vehicle 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

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

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.[13] 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 $H_0 : \hat{\beta}^{\text{LLM}} = \hat{\beta}$, 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 adoption). Values of β close to zero indicate that vehicles are being adopted as fast as 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 was 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 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.

13. 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.

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 detailed here and validated using historical vehicle registration tables drastically reduces critical parsing errors compared to standard tools and 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 economists, 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 becomes 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 Platt Boustan, and Adam Storeygard. 2025. "New Data and Insights in Regional and Urban Economics."

Arora, Abhishek, Emily Silcock, Melissa Dell, and Leander Heldring. 2024. "Contrastive Entity Coreference and Disambiguation for Historical Texts." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 6174–6186. Miami, Florida, USA: Association for Computational Linguistics, November.

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.*

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." *The 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.

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.” *The 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.” *The Journal of Transport History* 27 (1): 71–92.

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].

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.

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 Cconomic Growth.” *The 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.

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.

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.

Stuart, Bryan A., and Evan J. Taylor. 2021. "The Effect of Social Connectedness on Crime: Evidence from the Great Migration." *The 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].

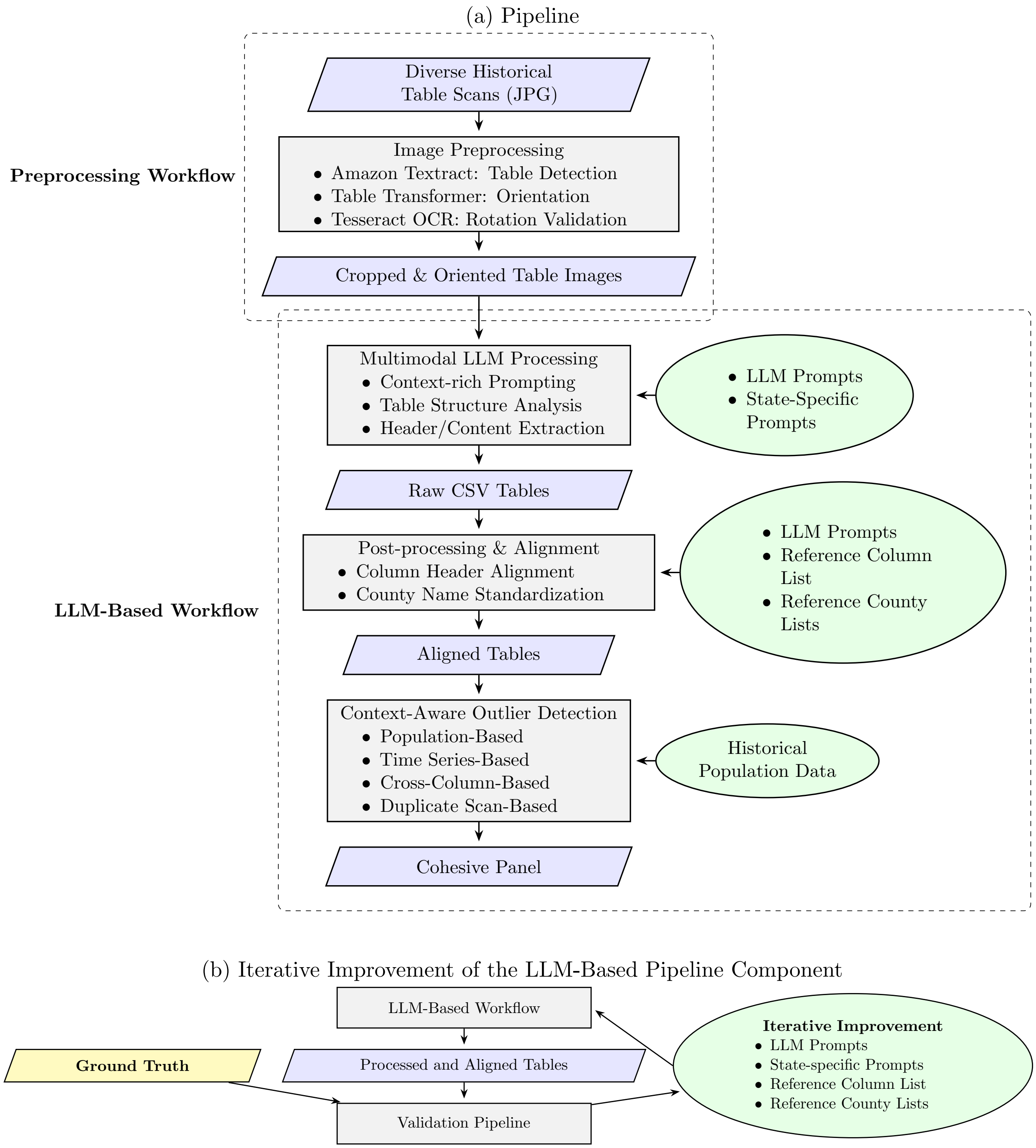


Figure 1: Historical Table Processing Pipeline

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).

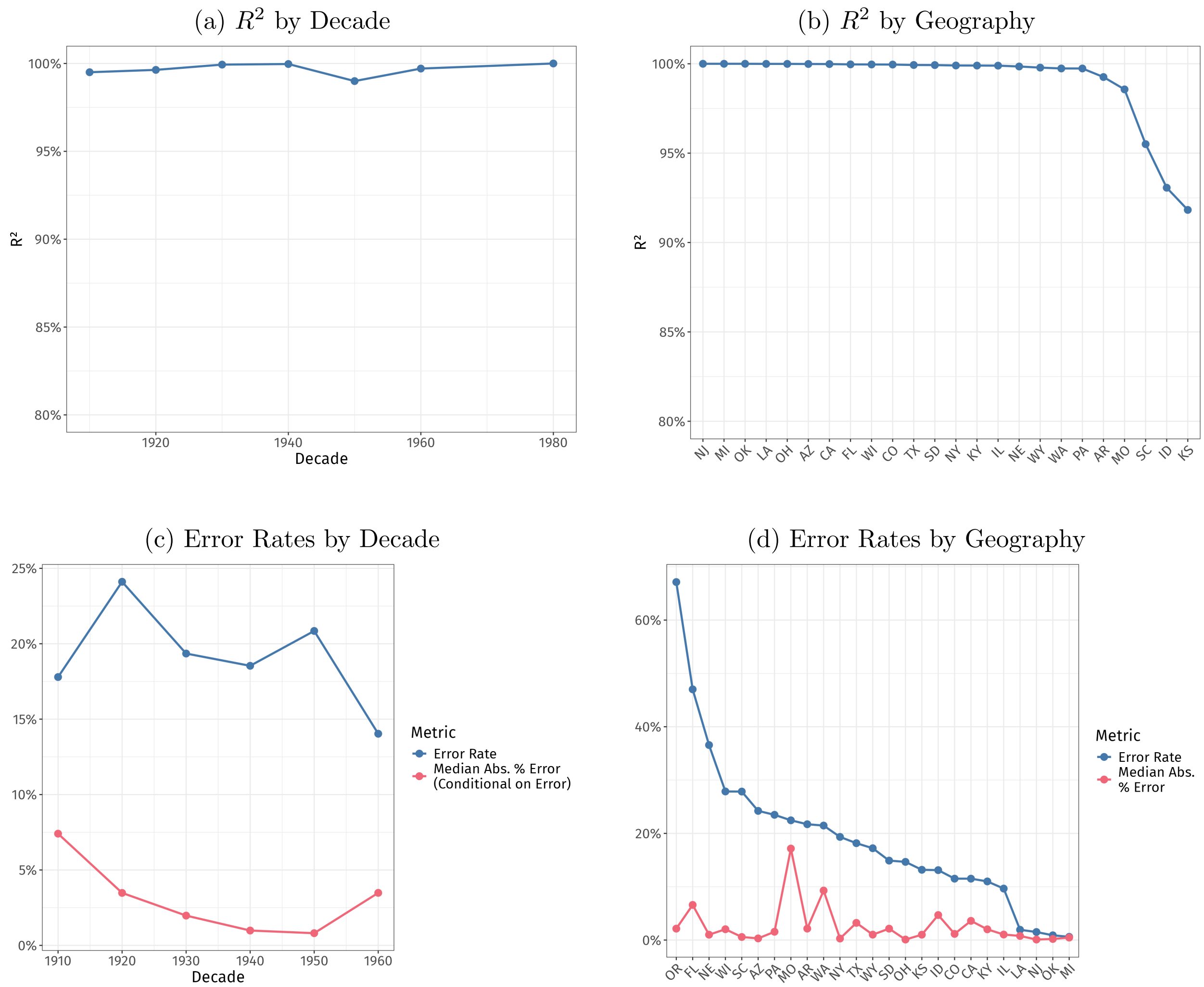


Figure 2: Performance Detail by Time and Geography

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.

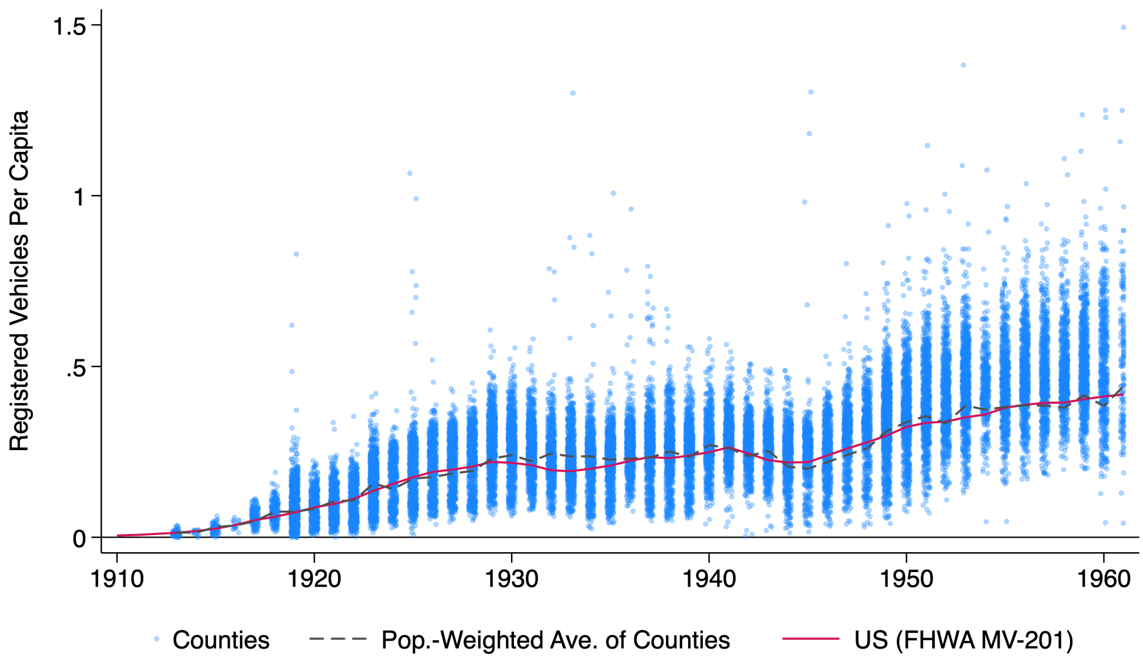


Figure 3: County-Level Vehicle Adoption Rates by Year, 1910–1961 (Cohesive Panel)

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} (1)	y_{cst} (2)	y_{cst}^{LLM} (3)	y_{cst} (4)	y_{cst}^{LLM} (5)	y_{cst} (6)	y_{cst}^{LLM} (7)	y_{cst} (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 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

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
Berricn	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.

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.

A.4 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.

Table A2: Performance Metrics (Holdout Sample)

(a) Overall Performance Metrics

	Model		
Metric	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

	Model		
Metric	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.

- *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 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.

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.

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.5 Determining the Size of 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. 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. 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.

A.6 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.

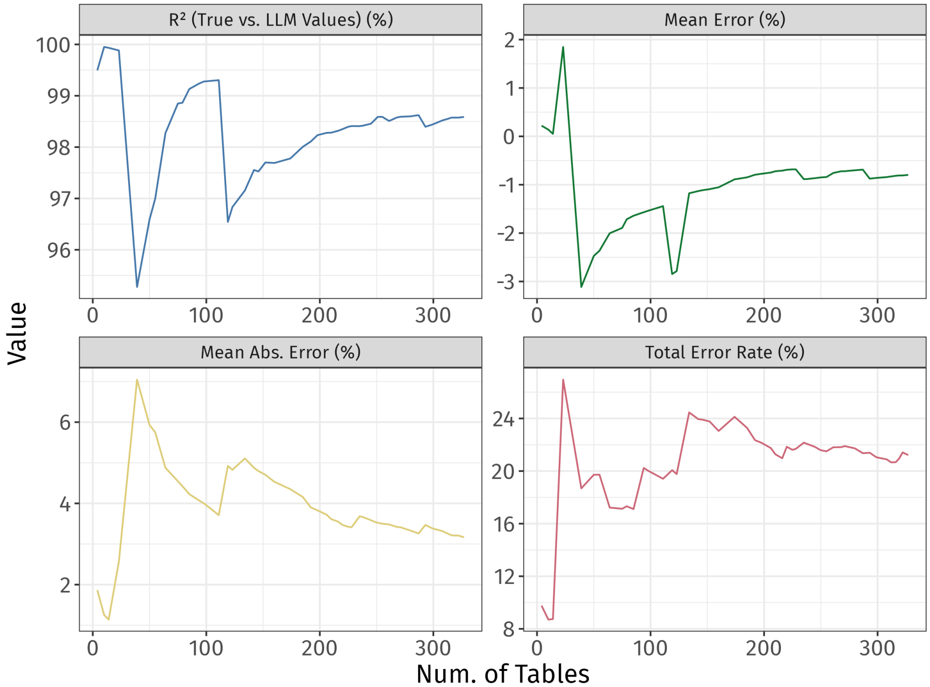
3. Among duplicates, preserve the cell(s) that returned values from a greater number LLM models.

4. For duplicates which return values from multiple LLMs, preserve the cell(s) for which the values across LLMs match.

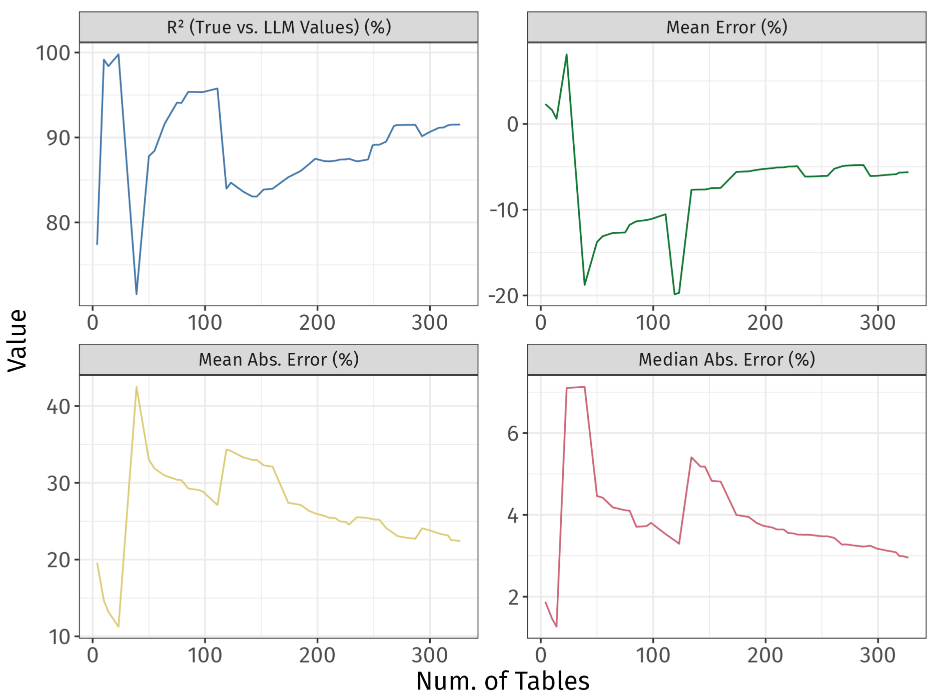
5. For duplicates that can be aggregated to a state-level total, preserve the cell(s) that contributes to the most accurate state-level total.

6. Among duplicates, preserve the cell(s) that is closest to the state-level per capita vehicle adoption rate.

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).

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).