The American Medical Directory Panel: Comparing AI-Assisted and LLM Transcription with Manual Data Entry at Scale*

Christian M. Dahl[a] Sophie Li[a] Rasmus Emil Nyholm[a]
Maja U. Pedersen[a] Anthony Wray[a†]

[a] University of Southern Denmark

November 17, 2026

How closely does machine transcription agree with manual double-entry data? We digitize the 1912 and 1921 editions of the American Medical Directory, a national register of US physicians with roughly 140,000 and 150,000 entries in the two volumes, and compare five machine-extraction workflows with a full-count double-entered manual transcription. Like many historical directories, the AMD is semi-structured rather than tabular, so errors arise from misreading the layout as well as the text. A script that parses commercial OCR text using rules written by coding agents recovers about 97 percent of manually transcribed records with similar precision, more than a multimodal language model workflow (gemini-2.0-flash) that reads the page images and the OCR text. When the two methods fail, they fail on partly overlapping sets of records, so reconciling their outputs raises both recall and precision to about 98 percent. An open-weight LLM pipeline run on a local GPU server falls between the parser and the multimodal workflow in recall, and replacing the commercial OCR model with an open-source alternative lowers its recall by only about 2 percentage points. As the manual transcription's own error rate is unknown, these rates measure agreement rather than accuracy. Direct processing charges for the commercial OCR and multimodal workflow were about €0.03 per page, compared with €14.70 per page for manual transcription. We release the data and codebook under CC BY 4.0.

JEL: C81, I11, N31, N32

Keywords: American Medical Directory; physicians; historical directories; OCR; large language models; transcription; record linkage; United States, 1912, 1921

*We acknowledge support from the Carlsberg Foundation through a Research Infrastructure Grant (PI: Christian M. Dahl) and a Semper Ardens: Accelerate grant (PI: Anthony Wray), and from Independent Research Fund Denmark through a Research Project 1 grant (PI: Anthony Wray). We thank Michel Rasmus Jessen, Freja Vallentin Witt, OpenAI's ChatGPT Desktop and Codex, and Anthropic's Claude Code for research assistance, Paradigm Data Services for the manual transcription, and Harish Bindra for coordinating it. We thank Gregory Niemesh for sharing the digital images of the AMD volumes, and the librarians at Miami University (Oxford, Ohio) for photographing the volumes and pre-processing the images. We thank seminar participants at the University of Gothenburg and the Boston University Empirical Micro Reading Group, and participants at the sixth Conference of the European Society of Historical Demography in Bologna (2025), for helpful comments. The data and codebook are available on Zenodo (version 1.0.0, DOI 10.5281/zenodo.23945788) under a CC BY 4.0 license.

†Corresponding author. Email: wray@sam.sdu.dk.

1 Introduction

How closely does machine transcription agree with manual double-entry data? The time required to key and clean historical records limits both dataset size and the questions economic historians can investigate (Gutmann et al. 2018). Advances in Optical Character Recognition (OCR) and large language models (LLMs) offer the potential to reduce the cost of digitizing historical sources, making large-scale transcription increasingly feasible (Dell 2025; Ferrara 2026). However, the benefits of automation depend on the accuracy of the resulting data. While automated transcription can reduce data-entry costs, assessing the quality of transcription still requires manual human review and raises new questions about validation. In particular, small character-level error rates can translate into substantial disagreement across records containing many fields (Correia and Luck 2023). Evaluating transcription accuracy therefore becomes critical.

Such evaluation raises a practical problem. Researchers may benchmark machine-generated data against a human-created reference dataset treated as ground truth, yet the reference dataset may itself contain transcription errors. Measures of agreement with a reference dataset therefore conflate machine transcription errors with errors in the benchmark, so agreement with a manual transcription does not necessarily establish accuracy (Dell and Rambachan 2026). This makes it difficult to determine which extraction strategy is most accurate. As AI-assisted digitization becomes more common, distinguishing agreement from accuracy becomes increasingly important for researchers releasing and using machine-generated historical datasets.

We study this question using the 1912 and 1921 editions of the *American Medical Directory* (AMD), a nationwide register of physicians in the United States published by the American Medical Association. The AMD records physician names, locations, medical schools, graduation years, licensing year, and professional credentials, making it one of the largest and most comprehensive individual-level sources on the early twentieth-century American medical profession. Despite its research potential, the directory remains difficult to use at scale because its information is embedded within compact, semi-structured entries rather than structured tables.

The AMD belongs to a broad class of historical sources that is common but difficult to digitize:

directories and registers that list individuals under place or category headings in compact, abbreviated entries, such as city and trade directories, professional registers, and membership rolls. Unlike tables, such sources have no fixed rows or columns. Entries in the AMD are arranged beneath town headings and contain numerous abbreviations, irregular formatting conventions, and nested record components. Errors can therefore arise not only from misreading individual fields but also from failing to identify record boundaries or town headings. Because a town heading determines the location assigned to every physician listed beneath it, a single error can affect many subsequent records. The AMD therefore allows us to distinguish errors in reading individual fields from errors in reconstructing the structure of the directory itself, a distinction that does not arise in the tabular sources studied in recent work (Bäcker-Peral et al. 2026).

To examine both kinds of error, we compare five extraction workflows against a large double-entered manual transcription of both editions, using the same pages, record-linkage rule, and measures for each. Precision is the share of a workflow's extracted records that link to a record in the manual transcription, and recall is the share of transcribed records that the workflow recovers; we compute both for whole records and for individual fields. Three workflows start from the same commercial OCR: a multimodal large language model that reads page images and OCR text, a deterministic parser of OCR text whose rules were written with coding agents, and a multi-stage open-weight LLM pipeline that reads OCR text alone. A fourth applies the same open-weight pipeline to OCR from an open-source model, which changes the OCR step (the engine and its column segmentation) and nothing downstream. The fifth, the combined table, reconciles the multimodal LLM and parser outputs field by field and forms the basis of the public release. Our data release includes 138,800 and 148,040 rows from the two editions, including cross-reference lines, covering 48 states and the District of Columbia. Because the case for automation rests on cost, we also compare the direct processing charges of the commercial OCR and MLLM workflow with those of the manual transcription.

Three results stand out. First, among the four single workflows, the deterministic parser has the highest recall in both editions. Its record-level precision and recall are 96.72 and 97.45 percent

in 1921 and 97.54 and 97.36 percent in 1912, higher on both measures than the multimodal LLM in both editions (94.11 and 94.13 percent in 1921; 94.89 and 94.42 percent in 1912) and higher on recall than the open-weight LLM pipeline. Replacing the commercial OCR with open-source OCR in that pipeline lowers recall by 2.4 percentage points in 1921 and 2.3 in 1912. Second, when the methods fail, they fail on partly overlapping sets of records. The combined table recovers more reference records than either input alone, with precision and recall of 98.04 and 98.28 percent in 1921 and 97.72 and 97.80 percent in 1912. Among cells where only the parser agrees with the transcription, most for birth and graduation year are entries the MLLM workflow did not link, while most for license year and medical school are different values on linked records. The widest gaps between methods are in county in both editions and in city in 1921. Third, processing charges for the commercial OCR and MLLM workflow are far below the price of manual transcription: an estimated €0.03 per page, against €14.70 per page for double-entered transcription. This comparison excludes engineering, review, and model-development time. Even so, digitization projects previously considered prohibitively expensive may now be feasible.

These figures should be read in light of the problem raised above: agreement with a reference dataset is not the same as accuracy. Our reference is a double-entered transcription whose own error rate is unknown, so the reported measures quantify agreement rather than scan-verified accuracy. Entries missed by both the machine and the transcribers are unobserved, and three of the workflows share the same OCR, so agreement among them is weaker evidence than agreement among independent readings would be. We therefore view the paper as a methodological benchmark of alternative extraction strategies and a description of the released data. A subsequent study will adjudicate disagreements directly against the page images and estimate accuracy using an independently verified gold-standard sample.

This paper contributes to a growing literature on large-scale historical transcription. Existing workflows typically combine commercial OCR with pre- and post-processing to transform scanned sources into structured datasets (Correia and Luck 2023, see Beach and Hanlon 2023 for newspapers). Other projects have developed specialized layout-detection and OCR pipelines for large

newspaper collections (Carlson et al. 2024), tabular documents (Dahl et al. 2023), and historical administrative directories (Fleischhacker et al. 2025); Dell (2025) surveys these approaches. More recently, large language models have been used to extract structured records directly from page images. Bäcker-Peral et al. (2026) find high agreement with a human-transcribed gold standard for historical county tables at substantially lower cost, with regression estimates of the same sign and significance, and Ferrara (2026) outlines a workflow that holds hand-coded validation data out of prompt development and model selection. We add to this literature in three ways. First, we compare multimodal, text-based LLM, and rule-based extraction on a common source and reference dataset, and we vary the OCR engine while holding the downstream pipeline fixed. Second, we study a directory-style source whose difficulty lies in its structure as much as in its characters, and we report which records and fields the methods fail to match. Third, we set out what agreement measures can and cannot establish when both machine-generated and human-generated transcriptions contain errors.

We also add to work that digitizes the AMD itself. Previous studies rely on hand-keyed samples for particular states and editions (Moehling et al. 2020), vendor double-entry transcriptions of North Carolina (Hollingsworth et al. 2024) and of the 1906 edition for Southern and adjacent states (Chrisinger 2024), and OCR extractions of single editions checked against printed physician counts (Smith 2017) or published state totals (Alsan and Neberai 2026). We provide national transcriptions of two editions validated against a double-entered reference. Appendix A summarizes earlier AMD datasets and their reported validation procedures.

Section 2 describes the directory. Section 3 describes the page images, the extraction workflows, and the public release. Section 4 describes the validation samples, the record-linkage rule, and the agreement measures, and Section 5 reports results for both editions. Section 6 reports robustness checks, and Section 7 discusses limitations and concludes with the planned adjudication against the page images. Appendix B documents the comparison protocol.

2 Background and Data: the American Medical Directory

The American Medical Directory (AMD) is a register of legally qualified and practicing physicians in the United States and Canada published every two to three years by the American Medical Association (AMA) starting in 1906 (American Medical Association 1912, 1921). The AMA compiled the AMD volumes from an index card collection that was updated daily using reports from licensing boards and lists of newly graduated physicians (American Medical Association 1912).[1] While the completeness of the AMD throughout its publication run remains an open question, the earliest editions may have omitted some physicians.[2] Among sources on the medical profession during the early-twentieth century, the directory uniquely captures individual physicians and their characteristics with nationwide coverage, and thus remains an invaluable resource for studying medical education and licensing, physician labor markets, and the professionalization of medical care. Appendix A provides an overview of prior hand-keyed transcriptions of AMD volumes and samples, and how these have been used in previous research.

Figure 1 shows a sample page of physician entries from the 1921 AMD volume, on the left, and closeups of two individual physician entries and one town heading with explanations of each field, on the right (American Medical Association 1921). The physician directory represents one section in a vast volume that includes listings of medical officers, medical schools, medical libraries, and hospitals. Each volume covers physicians in every state in the continental United States and the District of Columbia, as well as the US territories and Canada. The directory is organized hierarchically by state and town; within each state, towns are listed alphabetically. Each location heading reports the town name, its population, and county, and the physicians associated with that location are listed beneath it in alphabetical order. In some cases, a physician is listed in full under one town and appears elsewhere as a cross-reference.

The compact physician entries report medical education, credentials, professional affiliations,

[1] To compile the first edition, the AMA drew on records from state licensing boards, medical schools, medical societies, and a direct mail canvass of physicians conducted in 1905 and 1906 (Journal of the American Medical Association 1905).

[2] The first edition of 1906 likely undercounts Black physicians (Chrisinger 2024). In a few cases in each volume, an entry acknowledges that a practitioner's qualifications could not be independently verified.

and details about the physician's practice. A typical record follows a fixed order and format: the physician's name, sect letter, two-digit birth year, medical school code and two-digit graduation year, two-digit license year, membership symbol, specialty code, home address, and office address with office hours. Depending on the edition and page layout, the birth year and license year may be printed in parentheses. The exact conventions and entry order vary slightly across and within editions and do not always match the printed edition-specific key included in the directory. The abbreviations that appear in the AMD physician entries are explained in detail for the 1921 edition in Appendix C.

This project draws on physician entries from every state in the continental United States and the District of Columbia in the 1912 and 1921 editions of the American Medical Directory (American Medical Association 1912, 1921).[3] In this data note, we focus on a physician's name and location, the key dates (birth, graduation, and license year), and the medical school codes. Many of the remaining symbols require separate processing that lies outside the scope of this study.

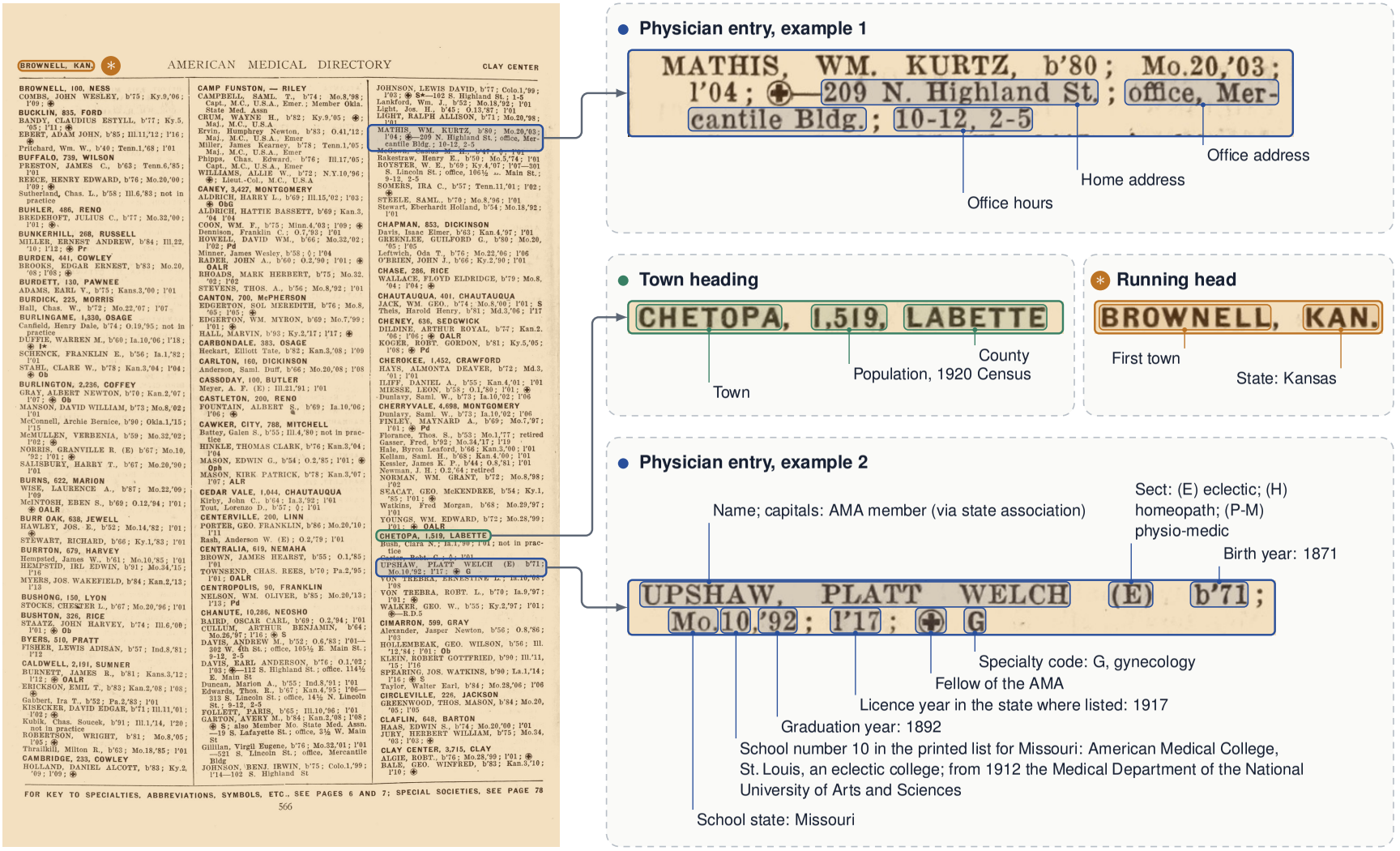
3 Digitizing the Directory: OCR and Structured Extraction Pipelines

This section sets out the challenges with structured output extraction from the AMD, describes the page images used as source inputs and the derived OCR outputs, explains the extraction methods, and summarizes the structure of the public data release.

The semi-structured text of the physician entries and their organizational structure make the AMD difficult to digitize. As the town heading determines the location assigned to every physician listed beneath it, errors in identifying a heading can affect many subsequent records even when the individual entries themselves are read correctly. However, the relevant town may appear in the page header, in a different column on the same page, or on a separate page altogether. Thus, the assignment of a physician entry to the correct location requires judgment over what constitutes an entry or a town header, in addition to preserving context across images or image segments.

[3] We exclude Canada and the U.S. territories from the public release as they lie outside our study population. We further exclude entries from the "Addendum" section that appears at the end of printed AMD volumes – containing corrections and late submissions – as we lack a manual transcription of these pages to assess precision and recall (see Section 4.1).

Figure 1. A page of the 1921 directory and the parts of an entry



Notes: Page 566 of the 1921 edition (Kansas), photographed from a library copy; the photograph is straightened, evenly lit and tinted, and each enlargement is straightened separately. The green box marks a town heading, which gives the town, its population and its county; every physician listed beneath it is assigned to that town. The orange box, marked with an asterisk, is the running head, which names the first town on the page and abbreviates the state (KAN., Kansas); town headings and entries do not repeat the state. The blue boxes mark two entries, enlarged on the right. Example 1, under Chanute, shows a home address, an office address and office hours; example 2, under Chetopa, shows the name, sect, birth year, school, graduation year, license year, Fellowship mark and specialty code. Labels follow the 1921 key (Appendix C) and the list of medical colleges printed in the same volume. Other pages of the 1921 volume print a second layout, with the birth year and license year in parentheses and the Fellowship mark after the birth year. Source: American Medical Association (1921).

These features make the AMD fundamentally different from tabular sources extracted at scale by other recent studies (Bäcker-Peral et al. 2026). Successful extraction therefore requires not only accurate character recognition but also correct identification of towns and record boundaries. For this reason, performance depends on recovering the structure of the document as well as the contents of individual entries.

3.1 Page images and OCR

Figure 2 provides an overview of the image pre-processing, OCR reading, structured output extraction, and validation workflows compared in this paper. Given that existing digitized versions of the AMD volumes available in the public domain consist of low-resolution color scans or poor-quality black and white scans, we obtained high-resolution digital images taken from physical copies of the AMD volumes as the source images for our extraction pipelines.[4] The pre-processing step consisted of cropping to ensure each image contained only the page itself, as well as a DPI adjustment.[5] These images represent a common input across all our extraction workflows.

3.2 Five machine extraction workflows

We evaluate five machine extraction workflows that each produce a structured dataset derived from page images of the directory. Table D.2 (Appendix D) summarizes their inputs and outputs. All extraction methods begin with the same underlying page images. Four of the workflows rely on OCR output generated by Azure Document Intelligence, a proprietary cloud API service that extracts both text and layout information (bounding box coordinates). We processed every page image with physician entries in the 1912 and 1921 editions during January 2026.

3.2.1 Multimodal language model workflow

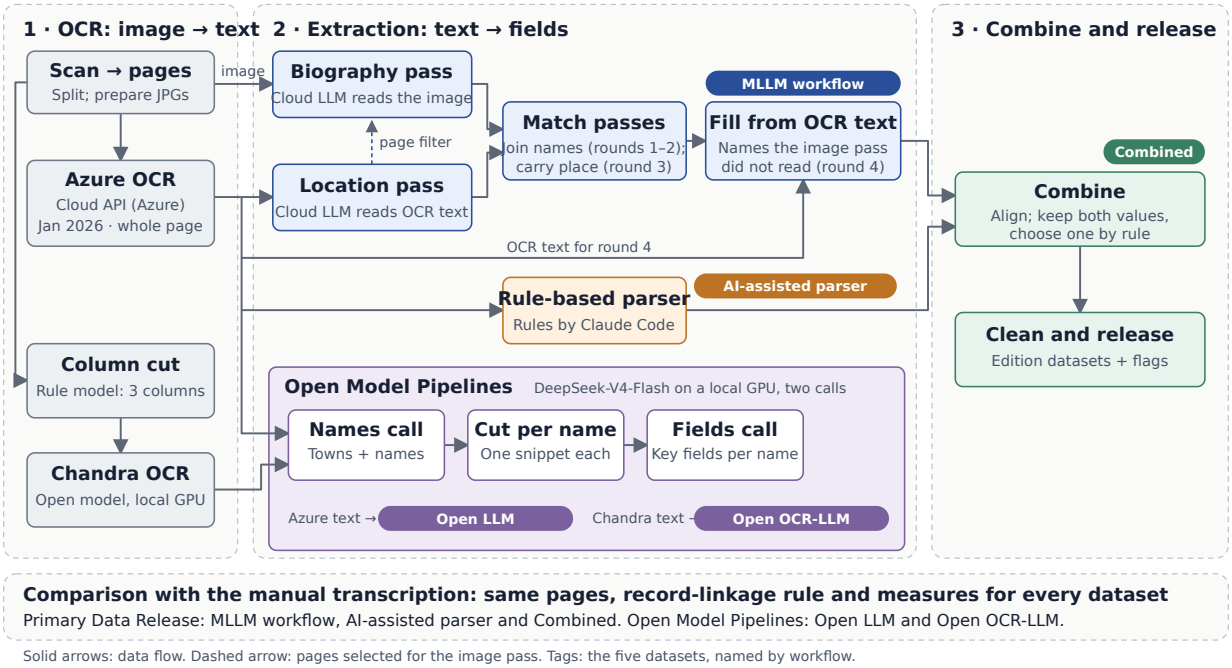
The multimodal language model (MLLM) workflow prompts the Google Gemini API and uses the gemini-2.0-flash model to extract information from the directory in two stages.[6] The first stage

[4]The images were taken by librarians at Miami University (Oxford, Ohio). We thank Gregory Niemesh for sharing the image files.

[5]As some pages were photographed multiple times, we de-duplicate the images prior to evaluation.

[6]Extraction was conducted using edition-specific prompts, and the resulting outputs from January 2026 for 1912 and February 2026 for 1921 were retained for evaluation. Because gemini-2.0-flash was discontinued on 1 June 2026, the workflow can no longer be rerun using the original model version.

Figure 2. From scans to the release tables



Notes: Solid arrows indicate data flow and the dashed arrow indicates page selection for image-based extraction. The manual transcription is used for validation and for selecting release-cleaning procedures, but it does not directly supply reconciled field values. Cross-edition linkage and record collapse occur downstream of the workflow examined in this paper. Source: project workflow documentation.

reads the OCR text generated by Azure Document Intelligence to identify the town and county headings as well as the physician names listed beneath them. The second stage reads the page image and extracts the biographical fields defined by the directory key. A carry-forward mechanism allows location information to be filled in across page and column boundaries (Appendix E describes the passes and matching rounds in detail). A deterministic linking algorithm then combines the list of towns and names from the first stage with the biographical information from the second into a single record-level dataset. The matching procedure blocks on page and name, while subsequent matching rounds allow for fuzzy matches on name and incorporate location information. A final stage uses the OCR text to fill missing biographical fields when the image-reading stage fails to extract them. This is the only stage in which OCR text contributes biographical information directly.

3.2.2 AI-assisted parser and combined workflow

As an alternative to prompting an MLLM via API calls, we prompted a coding agent to develop a deterministic script that parses the Azure OCR text. The AI-assisted parser follows a set of rule-based instructions to identify document sections, locate town headings, separate individual physician entries, and extract fields according to the conventions defined in the directory key. For every extracted field, it retains both the original OCR text and the cleaned value.

The combined workflow aligns the outputs of the MLLM and the AI-assisted parser at the record level to produce a combined data table. For each field, it retains both extracted values, records the rule used to determine the release value, and preserves the alternative value for reference. Both extracted values are kept for every field. Reconciliation uses only the two readings and fixed conventions (the medical-school code list, year-typo tables and county rules), not the manual transcription.

3.2.3 Open-source GPU-based OCR and LLM workflows

As one of our objectives is to evaluate the quality of data produced by open-source model in comparison to frontier models, our final two workflows use open weights and run on a GPU server. We start with the same Azure OCR text as the MLLM workflow, and replace the API calls to the gemini-2.0-flash model with a local deployment of DeepSeek-V4-Flash on a GPU server with temperature 0.01 to minimize variation across runs. The same prompts are used for both editions. Unlike the MLLM, the multi-stage open-LLM extraction pipeline operates only on OCR text rather than page images. The first stage identifies town headings and physician names on each page. Outputs are checked automatically for obvious failures, such as implausibly short name lists, and the model call is repeated when necessary. The OCR text is then divided into physician-specific snippets, which form the input for the second stage. A second model call extracts the biographical fields defined by the directory key.

Finally, we assess an end-to-end open source pipeline. The open OCR-LLM workflow uses OCR text generated by the open-source Chandra OCR model run locally on the GPU server, rather than the Azure OCR text. Pages are segmented into their printed columns using a table parsing

algorithm prior to processing by Chandra. These open LLM workflows produce two additional datasets that are evaluated against the same validation sample.

3.3 The public data release

The public data release through Zenodo (version 1.0.0, DOI 10.5281/zenodo.23945788) contains 138,800 rows for the 1912 edition and 148,040 rows for the 1921 edition, covering physicians in every state in the continental United States and the District of Columbia. Table 1 summarizes the public release tables. For each edition, the release includes a primary table, the combined workflow's output, which also holds the cross-reference lines. Each observation is identified by edition and record identifier. The outputs of the open-LLM and open-OCR-LLM workflows are provided as separate tables (see Section 5 and Appendix B for details). They are linked to the primary table through edition and record identifiers and include match-status information, but are distributed separately rather than merged into it. The agreement figures for the combined table in Section 5 describe the reconciled output with its county replaced by the cleaned county; Table G.3 reports the release table's columns as read and after cleaning. The release holds two cross-sections; it contains no identifier that links a physician across editions.

The first public release covers the state sections of the two volumes, corresponding to the material included in the manual transcription. The release tables include all fields evaluated in Section 5, with as-read and cleaned values for the years, medical school and county, and the workflow that found each entry. Additional extracted fields, including specialty, addresses and office hours, military information, printed marks (membership, fellowship, sect, and *col.*), and town populations, will remain unpublished until their validation is complete. These fields are documented in the codebook.

The release package also contains documentation, a codebook, a manifest identifying the included editions, checksums, citation information, and a CC BY 4.0 license. Both editions are in the public domain in the United States, which provides the basis for releasing the data tables. Appendix C lists the printed key elements and the release field each one feeds. The Zenodo record also includes a preferred citation.

Table 1. Public Data Release: Descriptive Statistics

	1912	1921
Listed physicians	137,303	145,608
Cross-reference lines	1,497	2,432
States and DC with coded rows	49	49
Physicians by Census region (%)		
Northeast	39,150 (28.5)	41,787 (28.7)
Midwest	47,238 (34.4)	48,185 (33.1)
South	39,195 (28.5)	41,292 (28.4)
West	11,657 (8.5)	14,284 (9.8)
State not resolved to a code	63 (0.0)	60 (0.0)
Share of physicians with a value (%)		
Birth year	73.8	89.6
Medical school	96.6	97.6
Graduation year	96.6	97.6
License year	78.0	85.2
City	100.0	100.0
County	99.8	99.5

Listed rows are printed entries, not distinct people; cross-reference lines are counted on their own row. Census regions are the Bureau's four regions over the 48 states and the District of Columbia; a row whose state is not resolved to a code is retained and counted on its own line. A field has a value when its cleaned column is non-missing; county and medical school are counted from the values as read. Source: descriptive statistics calculated from the public data release.

4 Validation

4.1 Double-entry manual transcription

We validate our machine-extraction workflows against a hand-keyed double-entry full-count transcription of the 1912 and 1921 AMD volumes. Each record was independently transcribed by two data entry workers and reconciled by a quality assurance checker. The manual transcription was completed during 2022, before LLM-based transcription tools were available.[7] The transcription includes physicians located in 48 US states and the District of Columbia, while excluding the addendum and address-unknown lists that appear at the back of the volumes. The resulting spreadsheets contain one row per printed entry and one column for each recorded field. We treat

[7]The first ChatGPT model released by OpenAI launched publicly on November 30, 2022 with limited capabilities and was based on GPT-3.5.

the compiled transcription as our reference dataset, which serves as the reference for evaluation.

4.2 Reference population for validation

Our reference population is the set of physician entries on printed volume pages for which we have a complete transcription. Table D.1 (Appendix D) summarizes the page coverage of the reference population in the 1912 and 1921 AMD volumes. For 1921, the manually transcribed reference dataset covers 1,233 pages in the printed directory and 145,064 cleaned rows. One complication with the manual transcription is that the data entry workers read from lower-quality scans of the 1912 and 1921 AMD volumes, rather than the high-resolution digital images. Consequently, some scanned pages or column segments did not contain any legible content; transcribers flagged such pages or columns in the output. After excluding pages with incomplete reference data, around 10 to 16 percent of the total, we evaluate all machine extraction methods on a sample consisting of 1,036 pages and 124,167 cleaned rows (Section 4.4). For 1912, the manual transcription covers 1,006 pages while the reference population consists of 903 pages containing 123,726 cleaned rows after excluding incomplete pages.

4.3 Record linkage to the validation sample

In the absence of a unique record identifier, we apply a record linkage procedure to merge the reference population from the manual transcription to the output of each machine-extraction workflow. The goal is to identify which extracted records correspond to which transcribed records. The resulting links form the basis for the measures of record-level agreement.

Prior to linking, we standardize the given names, medical school codes, two-digit years, and county names in the manual transcription and the extraction outputs (See Appendix B for details). We match records using combinations of page number and physician name whenever these identifiers are unique in both datasets. Remaining ambiguous cases are resolved using state and city information, but only when these fields uniquely identify a match in both datasets. School codes and years are never used to establish a link. Records in the manual transcription that cannot be linked, and ties that cannot be resolved, remain unmatched and continue to count toward precision

and recall denominators. Linkage is performed on the full set of extracted records before the analysis frame is applied. A field value can agree only on a linked record; values on unlinked records stay in the field denominators and count as disagreements.

4.4 Precision and recall measures

The central question in this paper is how closely each machine extraction method agrees with the reference population from the manual transcription. We validate the extraction methods by evaluating performance at both the record and field level. For a method m, let O_m denote the number of extracted records, N the number of reference records (on pages with complete transcription), and L_m the number of linked records produced by the linkage procedure described in Section 4.3. Record-level precision is defined as $\text{Precision}_m = L_m/O_m$, that is, the share of extracted records that can be linked to a reference record. Record recall is defined as $\text{Recall}_m = L_m/N$, or the share of reference records recovered by the extraction method. A reference record is recovered when it receives a one-to-one link in the first or second pass: a page-and-name key that is unique on both sides, or, within ambiguous groups, a unique state–city pair (Section 4.3). Field-level measures are calculated similarly. Precision divides the number of agreeing linked values by the total number of scorable output values produced by a method, while recall divides the number of agreeing linked values by the number of scorable values in the reference dataset. Cells containing multiple years or multiple school codes contribute each individual value to the comparison. Unmatched records count as disagreements, while cases where both sides are missing are excluded.

These measures quantify agreement with the reference population used for validation rather than accuracy relative to the printed directory. The distinction is important because while double-entry transcription reduces keying errors, the reference data may itself contain errors. To illustrate, consider a set of aligned records. Suppose the extraction method is incorrect with probability e_m and the reference transcription is incorrect with probability e_r. If errors are independent and never produce the same incorrect value, the probability of agreement is $A_m = (1 - e_m)(1 - e_r)$. Even in this highly simplified case, observed agreement reflects both error rates. Recovering the extraction error rate requires information about the accuracy of the reference transcription. Shared

errors may inflate agreement, while correlated errors invalidate the independence assumption altogether. High agreement therefore does not necessarily imply high accuracy. This distinction between agreement and accuracy is central to the evaluation of machine-generated measurements (Dell and Rambachan 2026). Throughout the paper, the reported precision and recall statistics should therefore be interpreted as measures of agreement with the validation sample – the share of transcribed entries recovered by an extraction method rather than the share of all printed physicians recovered from the original volumes. Estimating scan-verified accuracy would require an independent transcription of the same pages from which to adjudicate against the original page image and estimate the error rate of the manual transcription, which is planned for future work.

5 Results

This section reports the main results for the rate of agreement between the machine extraction methods and the cleaned reference population used for validation. The measures are reported for the set of pages in each edition with a complete manual transcription (903 pages in 1912 and 1,036 pages in 1921). They quantify agreement with the reference transcription rather than scan-verified accuracy. All headline comparisons are based on the common precision and recall measures defined in Section 4.4. Record-level agreement is reported in Table 2, while field-level agreement is summarized in Figure 3.

5.1 Record-level agreement

Table 2 provides the main record-level comparison across extraction methods. Precision measures the share of output records that can be linked to a transcribed record, while recall measures the share of transcribed records recovered by a method. We focus on comparisons that show how agreement changes when individual components of the extraction process are altered.

Overall performance: Three findings stand out. First, the AI-assisted parser consistently outperforms the MLLM workflow. Second, combining the MLLM and parser outputs produces the highest overall recall. Third, agreement is broadly similar across the 1912 and 1921 editions.

AI-assisted parser versus MLLM workflow: The clearest contrast is between the determinis-

Table 2. Record agreement by method, 1912 and 1921

	1912: 903 pages, 123,726 reference rows				1921: 1,036 pages, 124,167 reference rows			
Method	Output rows	Links	Prec. (%)	Recall (%)	Output rows	Links	Prec. (%)	Recall (%)
Primary table and its component workflows								
MLLM	123,108	116,817	94.89	94.42	124,195	116,883	94.11	94.13
Parser	123,496	120,462	97.54	97.36	125,101	120,998	96.72	97.45
Combined	123,829	121,010	97.72	97.80	124,477	122,032	98.04	98.28
Open model pipelines								
Open LLM	121,370	118,613	97.73	95.87	122,915	118,070	96.06	95.09
Open OCR-LLM	119,991	115,780	96.49	93.58	121,295	115,136	94.92	92.73

This table reports record-level precision and recall for each of the five machine extraction workflows and each AMD volume (1912 and 1921) compared to the manual transcription. The reference population for validation is the set of pages with complete manual transcription coverage. Measures reflect agreement with the cleaned validation sample. Recall uses the cleaned reference-row count, while precision uses each method's output-row count. All workflows count printed cross-reference lines as records, in both denominators and with the same linkage key.
Primary table and its component workflows: multimodal language model (MLLM), gemini-2.0-flash on the page image and the Azure OCR text; AI-assisted parser, a deterministic parser on the Azure OCR text; Combined, the parser and MLLM outputs merged field by field.
Open model pipelines: Open LLM, a multi-stage LLM pipeline (DeepSeek-V4-Flash) on the Azure OCR text; Open OCR-LLM, the same pipeline on OCR text from an open-source model (Chandra).

tic AI-assisted parser and the MLLM workflow. In both editions, the parser links more reference records and does so with higher precision. For 1921, the parser achieves precision and recall of 96.72/97.45%, compared with 94.11/94.13% for the MLLM workflow. The same pattern appears in 1912, where the parser reaches 97.54/97.36%, compared with 94.89/94.42%. The precision advantage is between two and three percentage points in both editions, while recall is 3.31 percentage points higher in 1921 and roughly three percentage points higher in 1912. For this source, the deterministic parser achieves both higher precision and higher recall than the multimodal extraction workflow.

Image-based and text-based extraction within the MLLM workflow: The MLLM workflow combines two stages implemented with gemini-2.0-flash. The location stage operates on OCR text, while the biography stage operates on the page image itself. Table D.4 shows that the two stages achieve remarkably similar linkage rates. The text-based stage links 91.64% of the transcribed rows at 95.44% precision, while the image-based stage links 92.68% at 95.64% precision. The differences are less than half a percentage point in precision and about one percentage point in

recall. These figures do not provide a pure comparison between text and image inputs. The two stages perform different tasks: the OCR-based stage identifies town headings and physician names, while the image-based stage extracts biographical information.

Combining extraction methods: The combined table achieves the highest recall in both editions. In 1921 it reaches 98.04/98.28%, compared with 96.72/97.45% for the parser. In 1912 it reaches 97.72/97.80%, compared with 97.54/97.36%. The gains are modest: recall rises by 0.83 percentage points in 1921 and 0.44 in 1912, and precision by 1.32 and 0.18. The two methods therefore fail on partly overlapping sets of records, although most of the reference records that the parser does not link remain unlinked in the combined table. The largest field-level change is in county (Section 5.2), where most of the gain comes from a county-cleaning rule that we selected by comparison with the transcription, not from reconciliation.

Open LLM versus AI-assisted parser: The comparison between the open LLM workflow and the AI-assisted parser contrasts two OCR-based extraction workflows: one based on a multi-stage LLM pipeline and one based on deterministic parsing rules. In 1921, the open LLM workflow achieves precision 0.66 percentage points below the parser and recall 2.36 percentage points lower. In 1912, precision is about 0.19 percentage points higher and recall is about one and a half percentage points below the parser (95.87% compared with 97.36%). This comparison should be interpreted descriptively rather than causally. The two workflows differ not only in their extraction models but also in prompting strategy and document handling.

Changing the OCR source: The open OCR-LLM workflow applies the same code, model, and prompts as the open LLM workflow, but its OCR comes from an open-source model run on pages that are first segmented into columns. The contrast therefore changes the OCR step, the engine and the column segmentation together, and nothing downstream. Recall falls by 2.4 percentage points in 1921 and 2.3 in 1912, and precision by 1.1 and 1.2 (Table 2).

Comparisons across editions: Precision and recall are broadly similar across the 1912 and 1921 editions, generally within about a point of the other edition for the released workflows. The largest exception is the open LLM workflow, whose precision is about 1.7 percentage points higher

in 1912 (97.73%) than in 1921 (96.06%). The comparison does not allow us to determine whether this difference arises from characteristics of the page images themselves or from edition-specific processing decisions. The editions differ in their printed keys and, for the MLLM workflow, in the prompts. The open model pipelines use the same prompts for both editions, so their difference across editions is not a prompt effect.

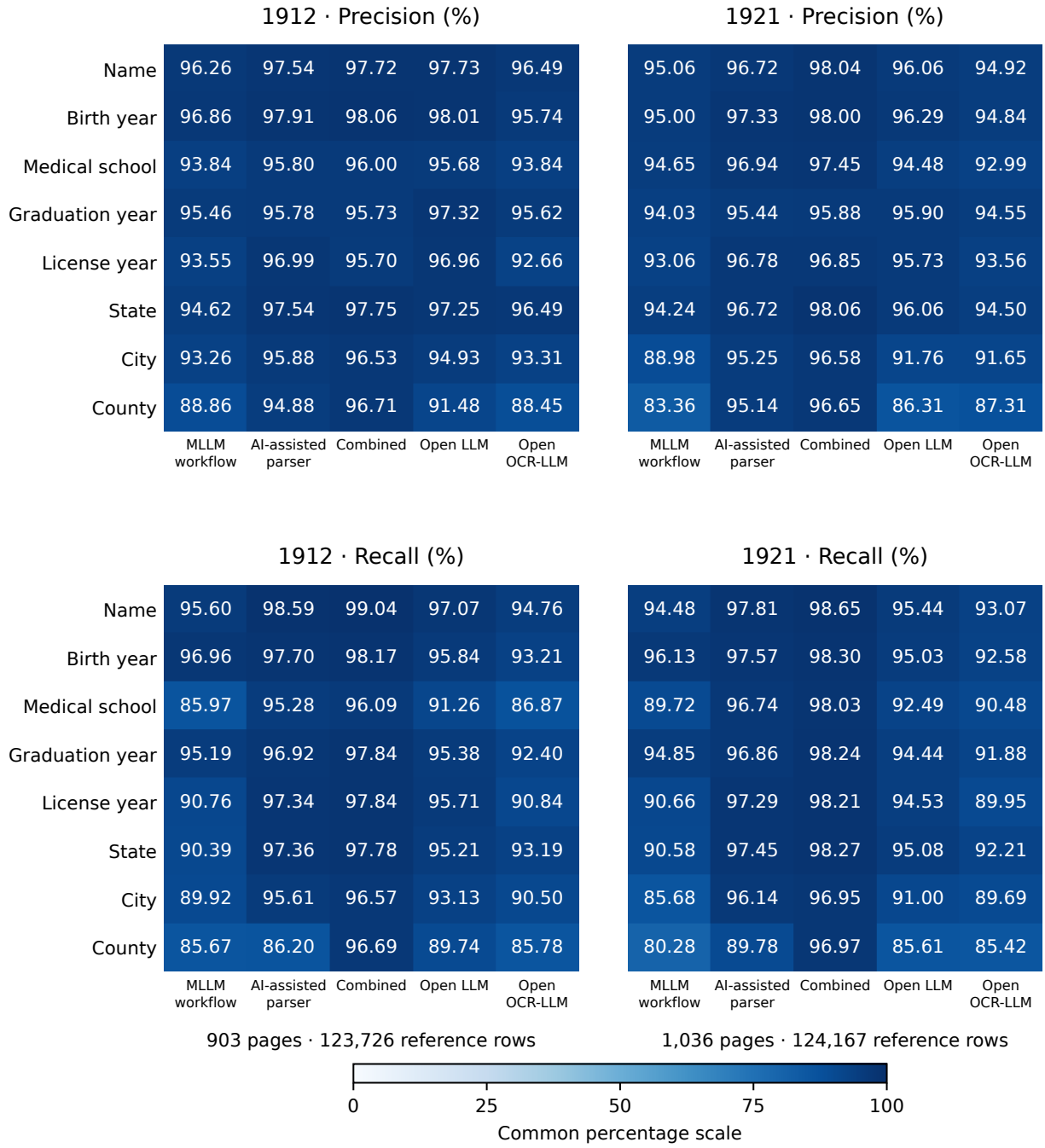
5.2 Field-level agreement

Record-level agreement masks important differences across fields. Figure 3 therefore examines agreement separately for names, locations, medical school codes, and other biographical characteristics. Exact percentages are reported in Table G.1 (Appendix G).

The broad pattern mirrors the record-level results. Across most biographical fields, the AI-assisted parser and the combined table achieve the highest levels of agreement with the reference transcription. In 1921, agreement for these methods ranges between 95 to 98 percent across the evaluated biographical fields. The corresponding range is 90 to 96 percent for the MLLM workflow and 92 to 96 percent for the open LLM workflow. For the open LLM workflow, medical school is the weakest of the four biographical fields, with recall of 92.49%. The same pattern appears in 1912, although field-level agreement varies somewhat more across fields. Medical school remains one of the most challenging fields for all methods and is again the weakest of the four biographical fields for the open LLM workflow (91.26%).

The largest differences across methods occur in place information. City agreement reaches 95 to 97 percent for the AI-assisted parser and the combined table in 1921, compared with 86 to 89 percent for the MLLM workflow. For 1912, the corresponding ranges are 96 to 97 and 90 to 93 percent. County shows the largest gain for the combined table, but most of it comes from county cleaning. In 1921, county precision and recall are 95.14 / 89.78 for the parser and 96.65 / 96.97 for the combined table. In 1912, the corresponding values are 94.88 / 86.20 and 96.71 / 96.69. The combined table's county is scored after the county-cleaning rule. As read, its 1921 figures are 90.90 / 90.96 (Table G.3). Reconciliation alone therefore raises county recall over the parser by about one point and lowers precision, and the cleaning rule accounts for the rest.

Figure 3. Field agreement by method, 1912 and 1921



Notes: Agreement with the cleaned validation sample; scan-verified accuracy not assessed. Same frames, links and scorer as Table 2. Precision divides agreeing linked values by all scorable output values; recall divides them by all scorable reference values. Unmatched values count as failures; missing/missing is excluded. Both editions and measures use the same 0–100% color scale, with percentages rounded to two decimals. Name and, for second-pass links, state/city agreement are linkage-conditioned. Years are two-digit tokens, school codes syntactically normalized, and places normalized for formatting. The combined table's county uses the checked county-cleaning rule on fixed links; its other fields use the reconciled reading. Table G.1 retains the exact displayed percentages; Table G.2 gives the 1921 integer counts.

5.3 Sources of disagreement

Understanding why methods disagree is as important as measuring how often they disagree. In many cases, disagreements arise not because a field value differs, but because a record is not linked: no record on the other side has the same page and name key, or the match is ambiguous. We currently have the most detailed diagnostic information for the 1921 edition. For the combined table, recall losses fall into three main groups: 1,569 transcribed records whose page-and-name key has no corresponding extracted record, 107 records whose key could not be resolved during second-pass matching, and 459 records lacking a usable name or page identifier. Precision losses arise from 2,327 machine-generated records with no corresponding transcription record and 118 records that remain ambiguous after matching. The available diagnostics cannot determine whether these discrepancies originate in the extraction process or in the reference transcription itself. Resolving these cases requires direct inspection of the page images.

Table G.4 (Appendix G) counts the disagreements at the field level, and Table D.5 (Appendix D) reports recall separately for entries flagged by the transcribers as uncertain. Among observations where only one of the two component methods behind the combined table agrees with the reference transcription, the parser agrees more frequently for all four evaluated fields. For birth year and graduation year, most of this advantage reflects missing records in the MLLM output. Missing entries account for 1,798 of 1,893 disagreements in birth year and 1,958 of 3,031 disagreements in graduation year. For license year and medical school, missing records explain a smaller share of disagreements. They account for only 1,594 of 5,676 license-year discrepancies and 1,955 of 7,106 school-code discrepancies. In the remaining cases, both outputs contain a record, but one workflow extracts a different field value from the reference transcription. In Table D.5 (earlier parser, first-pass links), disagreement is also more common among entries that the transcribers themselves flagged as uncertain. For the MLLM workflow, the parser and the combined table, record linkage recall is lower for entries marked as uncertain in the reference transcription.

5.4 Cost

The commercial OCR and MLLM workflow cost an estimated €0.03 per page in direct processing charges (Gemini plus Azure OCR), roughly 440 times less than manual double-entry transcription, which cost €14.70 per page. These figures cover processing charges only and leave out engineering, review, model development, coding-agent assistance and the local computing used to run the open model pipelines, which also serves other project work. The machine outputs and the manual transcription do not cover exactly the same pages and fields, so this is not an end-to-end productivity comparison (Appendix H).

6 Robustness

The ordering of the AI-assisted parser and the MLLM workflow is robust to the choice of reference population, links, and page restrictions. Using all manually transcribed pages, including those the transcribers flagged as having incomplete coverage, moves recall by about 0.3 points or less in both editions and 1912 precision by at most 0.16 points. It lowers 1921 precision by about 1.6 points for every workflow, because machine rows for entries the transcribers skipped count as errors (Appendix I). Recall is lower for entries the transcribers flagged as uncertain in all three workflows (Table D.5, Appendix D). Using first-pass links only shifts 1921 record agreement by about 0.2 points or less (Appendix D). Dropping the pages used in parser development changes precision and recall by at most 0.23 percentage points for every workflow in both editions, and the ordering of workflows by recall stays the same (Appendix I).

7 Conclusion

How closely does machine transcription agree with double-entered manual transcription at full-count scale? This paper compares five machine extraction workflows against a double-entered full-count transcription of the 1912 and 1921 volumes of the American Medical Directory, holding roughly 140,000 and 150,000 entries of physician biographies, respectively. Machine extraction agrees closely with the double-entered manual transcription.

In 1921 the AI-assisted parser recovers 97.45% of the manually transcribed entries and the

MLLM workflow 94.13%; in 1912 the figures are 97.36% and 94.42%. Precision is similar, at 96.72% and 94.11% in 1921 and 97.54% and 94.89% in 1912. The open LLM workflow falls between the MLLM workflow and the parser in recall. The open LLM workflow reaches 95.09% recall in 1921 and 95.87% in 1912, and replacing the commercial OCR with open-source OCR in the open OCR-LLM workflow costs 2.4 percentage points of recall in 1921 and 2.3 in 1912. The practical lesson is to combine workflows. The MLLM workflow and the AI-assisted parser fail on partly overlapping sets of records, and the combined table reaches 98.28% recall in 1921 and 97.80% in 1912, above any single workflow. Direct processing charges for the commercial OCR and MLLM workflow are an estimated €0.03 per page, against €14.70 per page for manual transcription, roughly 440 times less before engineering, review and hardware.

As we have emphasized, the reported precision and recall measures quantify agreement with the reference population, and do not measure scan-verified accuracy. Entries missed by both the extraction methods and the manual transcription are unobserved, so recall is defined relative to the reference transcription rather than to the printed directories themselves. Several questions therefore remain beyond the scope of the present comparison, including the accuracy of the reference transcription itself. The extraction methods are also not fully independent. The MLLM workflow, the AI-assisted parser, the combined table and the open LLM workflow all rely on the same Azure OCR text of the underlying page images. Agreement among these methods therefore provides less evidence than would agreement among fully independent readings. The disagreements examined in Section 5.3 provide useful diagnostic information, but they do not establish which source is correct. Resolving such cases requires adjudication against the original page images.

Several extensions are planned. First, we will construct a scan-verified evaluation sample based on direct adjudication against the original page images, allowing accuracy to be assessed rather than inferred from agreement. Second, we will classify disagreements by extraction stage and apply the construct-versus-observation framework of Dell and Rambachan (2026) to selected variables. Third, we will evaluate additional extraction workflows designed to provide a more independent comparison.

References

Alsan, Marcella and Yousra Neberai (2026) "Why Doesn't the United States Have National Health Insurance? The Political Role of the American Medical Association," *Quarterly Journal of Economics*, Vol. 141, No. 3, pp. 2147–2204.

American Medical Association (1912) *American Medical Directory: A Register of Legally Qualified Physicians of the United States, Alaska, Canal Zone, Hawaii, Porto Rico, Philippine Islands, Canada and Newfoundland*, Chicago: American Medical Association, 3rd edition.

——— (1921) *American Medical Directory*, Chicago: American Medical Association, 7th edition.

Andrews, Brendon Patrick (2023) "Physician Quality and the Flexner Report of 1910," Working paper, University of Alberta.

Beach, Brian and W. Walker Hanlon (2023) "Historical Newspaper Data: A Researcher's Guide," *Explorations in Economic History*, Vol. 90, p. 101541.

Bäcker-Peral, Verónica, Vitaly Meursault, and Christopher Severen (2026) "Can LLMs Credibly Transform the Creation of Panel Data from Diverse Historical Tables?," arXiv preprint arXiv:2505.11599v2.

Carlson, Jacob, Tom Bryan, and Melissa Dell (2024) "Efficient OCR for Building a Diverse Digital History," arXiv preprint arXiv:2304.02737.

Catillon, Maryaline, David Cutler, and Thomas Getzen (2018) "Two Hundred Years of Health and Medical Care: The Importance of Medical Care for Life Expectancy Gains," Working Paper No. 25330, National Bureau of Economic Research.

Chrisinger, Benjamin W. (2023) "Deep Roots of Racial Inequalities in US Healthcare: The 1906 American Medical Directory," Dataset, version 2, University of Oxford Sustainable Digital Scholarship portal, `https://doi.org/10.25446/oxford.24065709.v2`.

——— (2024) "US County-Level Variation in Availability and Prevalence of Black Physicians in 1906," *JAMA Network Open*, Vol. 7, No. 5, p. e2410242.

Correia, Sergio and Stephan Luck (2023) "Digitizing Historical Balance Sheet Data: A Practitioner's Guide," *Explorations in Economic History*, Vol. 87, p. 101475.

Dahl, Christian M., Torben S. D. Johansen, Emil N. Sørensen, Christian E. Westermann, and Simon Wittrock (2023) "Applications of machine learning in tabular document digitisation," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, Vol. 56, No. 1, pp. 34–48.

Dell, Melissa (2025) "Deep Learning for Economists," *Journal of Economic Literature*, Vol. 63, No. 1, pp. 5–58.

Dell, Melissa and Ashesh Rambachan (2026) "The Measurement Revolution? Credible Measurement and Inference in the Age of AI," Working Paper No. 35744, National Bureau of Economic Research.

Ferrara, Andreas (2026) "A Practitioner's Guide to Using Large Language Models and Generative AI in Economic History," Discussion Paper No. 171/26, Rockwool Foundation Berlin (RFBerlin).

Fleischhacker, David, Roman Kern, and Wolfgang Göderle (2025) "Enhancing OCR in Historical Documents with Complex Layouts through Machine Learning," *International Journal on Digital Libraries*, Vol. 26, p. 3.

Gutmann, Myron P., Emily Klancher Merchant, and Evan Roberts (2018) ""Big Data" in Economic History," *The Journal of Economic History*, Vol. 78, No. 1, pp. 268–299.

Hollingsworth, Alex, Krzysztof Karbownik, Melissa A. Thomasson, and Anthony Wray (2024) "The Gift of a Lifetime: The Hospital, Modern Medicine, and Mortality," *American Economic Review*, Vol. 114, No. 7, pp. 2201–2238.

Journal of the American Medical Association (1905) "Biographical Card Index and Directory," *JAMA*, Vol. 45, No. 21, pp. 1574–1575.

McNamara, Robert L., Edward W. Hassinger, and John B. Mitchell (1955) "Supply of Physicians in Rural Missouri," Bulletin 651, University of Missouri College of Agriculture, Agricultural Experiment Station.

Moehling, Carolyn M., Gregory T. Niemesh, Melissa A. Thomasson, and Jaret Treber (2020) "Medical Education Reforms and the Origins of the Rural Physician Shortage," *Cliometrica*, Vol. 14, pp. 181–225.

Mountin, Joseph W., Elliott H. Pennell, and Virginia Nicolay (1942) "Location and Movement of Physicians, 1923 and 1938: General Observations," *Public Health Reports (1896–1970)*, Vol. 57, No. 37, pp. 1363–1375.

Mullally, Sasha and Katherine MacDonald (2017) "Call the Doctor? Understanding Health Service Trends in New Brunswick, Part I, 1918–1950," *Journal of New Brunswick Studies*, Vol. 8.

Schafer, Jr., James A. (2014) *The Business of Private Medical Practice: Doctors, Specialization, and Urban Change in Philadelphia, 1900–1940*, New Brunswick, NJ: Rutgers University Press.

Smith, Sean Morey (2017) "Digitizing Doctors: Methodologies for Creating a Database from Historical Directories of Physicians," *Medical History*, Vol. 61, No. 4, pp. 611–614.

Staines, Amber (2015) "The Effect of Medical Care on Infant Mortality in the United States in the Early 20th Century," M.A. thesis, Miami University.

Weisz, George (2006) "Regulating Specialists in National Medical Directories," in *Divide and Conquer: A Comparative History of Medical Specialization*, Oxford: Oxford University Press, pp. 87–104.

A Prior transcriptions of the AMD

Because successive editions repeat a recurring core of fields for the licensed profession, the AMD supports three kinds of study that few other historical sources allow: the supply and location of physicians in a place and its change over time (Moehling et al. 2020; Mountin et al. 1942); the training of physicians, through the medical school code and graduation year; and the careers of groups the directory marks, Black physicians through the *col.* label (Chrisinger 2024) and members and specialists through the membership marks and specialty codes (Alsan and Neberai 2026).

Table A.1 lists prior AMD transcriptions, including hand-keyed subsets and national OCR extractions of single editions, 1918 for Smith (2017) and 1950 for Alsan and Neberai (2026). Moehling et al. (2020) used hand transcriptions of four editions for four states to study physician labor markets. Our own earlier work used a vendor double-entry transcription of the AMD for North Carolina from every edition between 1912 and 1942 (Hollingsworth et al. 2024). Alsan and Neberai (2026) digitized and OCR'd the 1950 edition for membership and specialists and checked state totals against published aggregates. Other studies use a state, a city or sample (Staines 2015; Catillon et al. 2018; Mullally and MacDonald 2017; McNamara et al. 1955; Weisz 2006; Schafer 2014). Andrews (2023) uses AMD aggregates and a state directory, not AMD microdata.

The closest precedent to our validation samples is Chrisinger (2024), which used a double-entry manual transcription for eighteen Southern and adjacent states and the District of Columbia. The article analyzes 41,828 entries, including 632 physicians labeled *col.* and 113 unlabeled graduates of Black medical schools, and its dataset is publicly deposited (Chrisinger 2023). Smith (2017) provides a machine-transcription precedent with reported record checks. The 1918 prototype uses commercial OCR and two parsing programs, keeps 70 to 90 percent of single-line entries, varying by state, and reports more than 92 percent of manually checked entries as perfect. It also compares capture with printed counts for the largest cities.

A series of articles in *Public Health Reports* by Mountin, Pennell and coauthors examines physician location and movement between 1923 and 1938 (for example Mountin et al. 1942). Their titles and records show the subject but not how physicians were linked across editions.

Table A.1. Prior transcriptions of the American Medical Directory

Study	Editions	Coverage	Method	Validation reported	Public
Moehling et al. (2020)	1909, 1914, 1918, 1921	CA, MS, NY, NC	hand keying	—	—
Hollingsworth et al. (2024)	1912–1942	North Carolina	vendor double entry[a]	—	yes, deposited
Alsan and Neberai (2026)	1950 by OCR; earlier editions cited	national (membership, specialists)	OCR	state totals vs published aggregates	—
Smith (2017)	1918	national pilot	OCR, two parsers	printed per-city counts; 70 to 90% of single-line entries kept; >92% of checked entries perfect	—
Chrisinger (2024)	1906	18 Southern and adjacent states, DC	vendor double entry	likely undercount reported	yes, deposited
Mountin et al. (1942)	1923, 1938	national	Public Health Service tabulation	—	in print
This note	1912, 1921	48 states and DC	OCR; MLLM, AI-assisted parser, open LLM and open OCR-LLM workflows	agreement with a double-entered transcription on its completely transcribed pages	yes, Zenodo, CC BY

[a] Our own earlier work; the vendor reports double entry.
— : not examined for this note.

B Scoring protocol

This appendix collects the rules under which every figure in Section 5 was produced.

Record linkage. Each workflow is linked to the validation sample separately, with the same rule in both editions. The first pass matches printed-page and name keys that occur exactly once on each side. Within page-and-name groups that remain ambiguous, the second pass matches state–city pairs that are valid and unique on both sides. Birth, graduation and license years and the school code never select a link. Unmatched rows and unresolved ties remain in both denominators, and links are fixed over the whole edition before any page restriction, so name agreement is conditioned on the rule. The linkage key is the printed page and the name as a multiset of tokens: names are normalized (Unicode composition, case folding), split into alphanumeric tokens and sorted, with no fuzzy repair, and a blank name gives no key. Years and school codes are never keys. Table D.3 (Appendix D) gives the 1921 figures after the first pass alone.

Scorer. All headline tables use a common scorer. The scorer applies the same normalization to city and county on both sides before linking: it removes accents, the *St.*/*Mt.*/*Ft.* abbreviations, a trailing *County*, punctuation and spacing, and nothing else. Field precision divides agreeing linked values by all scorable output values and field recall by all scorable reference values; an unmatched row's values count as failures; a blank on both sides never counts as agreement; years are compared as two-digit tokens and school codes after the same syntactic normalization on both sides.

Cleaned reference. The scorer reads a *cleaned* version of the transcription, which normalizes school codes and two-digit years to the forms the key defines and does not touch names or places. Its largest effect is on the school code, where the transcribers' varied spellings of one code would otherwise count as disagreement. That difference reflects how the reference represents a value, not the extraction.

Frames. The analysis frame of an edition is the set of transcribed printed pages *without known incomplete coverage*: pages on which the compiled transcription records no ignored-row placeholder, including the pages used to tune the parser, the pages drawn into a sampling design and

the diagnostic pages (Appendix I reports the frame without them). For 1921 that is 1,036 pages and 124,167 cleaned rows out of 1,233 transcribed page keys, after dropping the 197 pages with an ignored-row placeholder. For 1912 it is 903 pages and 123,726 rows out of 1,006 page keys, after dropping the 103 pages with an ignored-row placeholder. Table D.1 (Appendix D) sets the two editions side by side; Table D.3 gives the 1921 figures on the broad frame.

Evaluated inputs. The MLLM workflow is the matched output of the stored biography and location passes; the 1921 model passes date from February 2026, and the 1912 output is the one used in the tables prepared for release. The AI-assisted parser is the parser's raw candidate list for each edition. The combined table is the reconciled output in each edition, with the checked cleaned county substituted for the as-read county. The paired county checks retain exactly the same links and all non-county fields. The open LLM workflow is open-model-pipeline output for 1921 (162,215 evaluated rows, of 162,947 output rows with 732 duplicate-page rows dropped) and 1912 (149,471 rows). The open OCR-LLM workflow is the output of the same open model pipeline, with the same model and prompts, on the open-source OCR text of each edition. The open LLM and open OCR-LLM outputs carry their own page identifiers, which a reviewed page bridge maps to printed pages in both editions. Table D.2 (Appendix D) lists the inputs of each workflow.

C The printed key for 1921

What an entry holds. The body is arranged by state, then by town, with the town's population and county on a heading line and the physicians beneath it in alphabetical order, in narrow columns. A physician listed in full under one town may be printed under another as a cross-reference line. Each entry packs its fields into a fixed order of abbreviations, decoded by an edition-specific printed key; the front-matter key is not in our scans, and we read it from an Internet Archive copy (American Medical Association 1921). Table C.1 lists the printed elements and the release field each feeds. The license year is the year of license in the state where the physician is listed, so a mover carries a new license year in the next edition; a missing year has several printed meanings, most with their own mark, and some states print no year by convention. The release tables carry the year only, not which reason the mark gave. The address element `R.D.`, the physician count, the town line and the cross-reference are layout elements rather than key definitions; we read their meanings from the entries themselves.

Table C.1. The 1921 printed key and release fields

Printed element	Meaning in 1921	Release field
Name in capitals	membership in the AMA through a constituent state association and component county society	withheld
AMA mark, letter-coded before OCR, after the license year[c]	Fellowship in the AMA; commissioned medical officers on active service are Fellows[a]	withheld
`(E)`, `(H)`, `(P-M)` after the name	eclectic, homeopath, physio-medic	withheld
`b'75`	born 1875	birth year, as read and cleaned
`Md.7,'09`	school key number (state abbreviation and number) and year of graduation	school code (cleaned to the key's form), graduation year
♢	information received does not show graduation from a medical college[a]	not captured
△	graduate of an institution not listed as a medical college by the Council on Medical Education and Hospitals[a]	not captured
`l'20`	year of license in the state where located	license year
`l`†	licensed, year not known	not captured
‡	information received does not show license in the state where located	not captured
circle under a cross	in practice in the state where located before passage of the present law, and licensed by filing a diploma or by years of practice	not captured
Specialty code (`S`, `Oph`, `Pd`, ...), *	self-reported interest; with the asterisk, practice claimed limited to that branch; printed on request only	withheld
`R.D.`, `R.D.1`	rural-delivery address element, never a specialty	withheld with the addresses
`(col.)` after the name	the Association's label for a Black physician[b]	withheld
`NUMBER OF PHYSICIANS, n` under a town line	count of physicians listed under that town	not released
Town line `TOWN, population, COUNTY`	population and county of the town	county, as read and cleaned; population withheld
`(See Chicago)`	cross-reference to the full entry under another town	cross-reference line in the release table
ADDENDA; index symbols	physicians received too late to classify; corrections; reports of deaths	not released

[a] Carried from the 1918 key or the 1921 addenda key; the 1921 front-matter key is not in our scans.

[b] Not defined in any key we have seen; the label appears only in entries, and its reading follows Chrisinger (2024).

[c] Or after the birth year on pages printed in the second layout.

D Linkage sensitivity and stage decomposition

Appendix B states the linkage rule, scorer, frames and inputs. Table D.1 gives the coverage of the validation samples (Section 4.1) and Table D.2 the five extraction workflows (Section 3.2).

Table D.1. Coverage of the validation samples, 1912 and 1921

	1912	1921
Printed-page keys transcribed (broad frame)	1,006	1,233
Cleaned rows, broad frame	137,501	145,064
Pages without known incomplete coverage (analysis frame)	903	1,036
Cleaned rows, analysis frame	123,726	124,167
Analysis pages seen while tuning the parser	93	116
Analysis pages in the frozen design sample	16	16
Analysis pages used in diagnostics	0	2
Fields transcribed	name, birth year, school, graduation year, license year, city, county, state, population	
Not transcribed	addenda; address-unknown list; count lines (inconsistently)	

Sources: comparison results and validation-sample documentation for each edition.

Table D.3 reports the 1921 record comparison of Section 5.1 under the first pass alone and after the second pass, and on the broad 1,233-page frame of every transcribed page key (145,064 cleaned rows), which includes the pages with known incomplete coverage. The broad frame lowers precision for every workflow: on a page the transcribers covered only in part, a machine row for a physician they skipped counts against the extraction. Table D.4 decomposes the 1921 MLLM workflow by stage on the analysis frame; Appendix E describes the stages. Table D.5 splits recall by the transcription's own flags, the evidence for Section 5.3.

Table D.2. The five extraction workflows

Workflow	Input	Model or rules	Fields produced	Evaluated output (1921)
MLLM	page image (biography); OCR text (location)	gemini-2.0-flash, prompts per edition; page matching; OCR enhancement	name, birth year, school, graduation year, license year, specialty, city, county, state	biography and location runs of February 2026; matching run of October 2026
Parser (AI-assisted)	OCR text	deterministic parser; same rules and dictionary in both editions	all of the above, plus population, membership and military flags, cross-reference lines	raw candidate list (Appendix B)
Combined	MLLM and AI-assisted parser	per-field reconciliation rules; release cleaning; 48-state rule	the parser's fields with chosen values and per-field provenance	reconciled output; checked cleaned county on fixed links
Open LLM	OCR text	multi-stage LLM extraction (open model pipeline)	name, birth year, school, graduation year, license year, city, county, state	1921: 162,215 evaluated rows (of 162,947 output rows, with 732 duplicate-page rows dropped)
Open OCR-LLM	OCR text of the page images by an open-source model	open model pipeline (same code, model and prompts)	as Open LLM	scores in Tables 2 and G.1

The MLLM workflow, the AI-assisted parser, the combined table and the open LLM workflow consume the same stored OCR output; the MLLM workflow also reads page images for biographical fields. The open OCR-LLM workflow consumes a separate OCR of the same page images and shares no OCR text with the others. Specialty and symbol flags are neither scored here nor included in the public release.

Table D.3. Record agreement by link pass and page frame, 1921 (precision / recall, %)

	Analysis frame, 1,036 pages			Analysis frame		Broad frame
Workflow	First-pass links	Second pass adds	Output rows	First pass	Second pass	Second pass
MLLM	116,602	281	124,195	93.89 / 93.91	94.11 / 94.13	92.55 / 93.96
Parser	120,923	75	125,101	96.66 / 97.39	96.72 / 97.45	95.11 / 97.44
Combined	121,949	83	124,477	97.97 / 98.21	98.04 / 98.28	96.43 / 98.32
Open LLM	117,987	83	122,915	95.99 / 95.02	96.06 / 95.09	94.44 / 95.22

Agreement with the cleaned validation sample; scan-verified accuracy not assessed. Broad frame: 1,233 page keys, 145,064 cleaned rows. Table 2 reports the open OCR-LLM scores on the analysis frame. Source: the common scorer's comparison results.

Table D.4. The MLLM workflow by stage, 1921

Stage	Precision (%)	Recall (%)
Location pass alone (OCR text)	95.44	91.64
Biography pass alone (image)	95.64	92.68
Matched, before OCR enhancement	94.76	92.77
Matched, after OCR enhancement (MLLM workflow)	94.11	94.13

Agreement with the cleaned validation sample; scan-verified accuracy not assessed. Analysis frame; second-pass links. The biography pass carries no geography and links on the first pass only. Source: the stage comparison.

Table D.5. Historical recall by the transcription's own uncertainty flags, 1921

Pages	Entries	Reference rows	MLLM	Parser	Combined
$\geqslant$5% flagged (37 pages)	flagged	726	60.88	60.47	61.29
	unflagged	3,763	89.90	90.70	91.10
$<$5% flagged (866 pages)	flagged	681	78.56	78.71	80.32
	unflagged	102,495	93.60	95.58	95.99

Agreement with the cleaned validation sample; scan-verified accuracy not assessed. Recall in percent. Historical inputs: the earlier parser and first-pass-only comparison, on the historical 1921 frame that predates the current analysis frame. These are not estimates for the parser in Table 2. These figures differ from Table 2; the open LLM and open OCR-LLM workflows are not reported. Precision is unavailable for these strata because an unmatched machine row has no stratum.

E The MLLM workflow in detail

This appendix describes the stored 1921 implementation of the MLLM workflow (Section 3.2.1): the two model passes, the checks on their output and the matching that joins them. Every call is to gemini-2.0-flash through the vendor's API with a prompt written for the 1921 edition; the stored runs date from February 2026 (Appendix F), and no call has been repeated since the model was shut down on 1 June 2026.

Location pass. The first pass reads the OCR text of a page, not the image. The prompt asks for structured JSON: the state, each town heading with its county and population, and the physician names printed beneath it. A page whose output contains neither a town heading nor a physician name is skipped as presumed non-entry material (an institution list, front or back matter); this rule can also exclude a physician page whose headings the pass fails to extract.

Biography pass. The second pass reads the page image. The model first transcribes the page with its layout into sections, then extracts from each section the physicians and the fields the key defines. A section that fails the schema check is sent back once for repair, and a section whose output overran the model's limit is split and read again. Pages and entries with none of the four key fields (birth year, school, graduation year, license year) are dropped at collection.

Matching. The matcher joins the two passes in four greedy, one-to-one rounds, so a name matches at most once: on the normalized name and its occurrence on the page; on the name alone by Jaro–Winkler similarity at or above 0.85; by carrying a location forward from the previous page to a biography that has none; and, last, by filling the fields of a name the location pass found but the image pass did not from the OCR text. That last step is the only one at which OCR text supplies biographical values to this workflow; it filled 2,958 rows in 1921 (2,632 on the analysis frame). Table D.4 (Appendix D) scores each stage on the 1921 analysis frame: each pass alone links 92 to 93 percent of the transcribed rows at about 95 to 96 percent precision, the two matched without the OCR step reach 92.77% at 94.76%, and the full extraction 94.13% at 94.11%. The OCR step adds 1,691 links and about 1.4 points of recall at a small cost in precision.

F Reproducibility statement

The comparisons use stored outputs; no OCR call, model call or extraction was made for this note. Appendix B describes the linkage rule, scorer, frames and inputs. This appendix describes what is released and the limits to reproducing the results.

Extraction and release. The stored inputs include the OCR text, the MLLM workflow's biography and location passes and their matched output, the parser candidates and the combined tables. For 1921, OCR was performed in January 2026 and the model passes in February 2026, using gemini-2.0-flash. For the combined table, county is taken from the separately checked county-cleaning rule on identical links; the other fields remain those of the reconciled output. The public dataset is identified by Zenodo version 1.0.0 and DOI 10.5281/zenodo.23945788. The descriptive statistics (Table 1) are calculated from the public release tables and can be reproduced from the public files.

Limits to reproduction. The model gemini-2.0-flash was shut down on 1 June 2026, so the MLLM extraction cannot be rerun through the vendor's API. The manual transcriptions are excluded from the public release, which prevents full public recomputation of the agreement results.

G Field tables by workflow

Table G.1 gives the percentages shown in Figure 3, for both editions.

Table G.1. Field agreement by workflow, 1912 and 1921 (precision / recall, %)

Field	MLLM	Parser	Combined	Open LLM	Open OCR-LLM
Panel A: 1912					
Name	96.26 / 95.60	97.54 / 98.59	97.72 / 99.04	97.73 / 97.07	96.49 / 94.76
Birth year	96.86 / 96.96	97.91 / 97.70	98.06 / 98.17	98.01 / 95.84	95.74 / 93.21
Medical school	93.84 / 85.97	95.80 / 95.28	96.00 / 96.09	95.68 / 91.26	93.84 / 86.87
Graduation year	95.46 / 95.19	95.78 / 96.92	95.73 / 97.84	97.32 / 95.38	95.62 / 92.40
License year	93.55 / 90.76	96.99 / 97.34	95.70 / 97.84	96.96 / 95.71	92.66 / 90.84
State	94.62 / 90.39	97.54 / 97.36	97.75 / 97.78	97.25 / 95.21	96.49 / 93.19
City	93.26 / 89.92	95.88 / 95.61	96.53 / 96.57	94.93 / 93.13	93.31 / 90.50
County	88.86 / 85.67	94.88 / 86.20	96.71 / 96.69	91.48 / 89.74	88.45 / 85.78
Panel B: 1921					
Name	95.06 / 94.48	96.72 / 97.81	98.04 / 98.65	96.06 / 95.44	94.92 / 93.07
Birth year	95.00 / 96.13	97.33 / 97.57	98.00 / 98.30	96.29 / 95.03	94.84 / 92.58
Medical school	94.65 / 89.72	96.94 / 96.74	97.45 / 98.03	94.48 / 92.49	92.99 / 90.48
Graduation year	94.03 / 94.85	95.44 / 96.86	95.88 / 98.24	95.90 / 94.44	94.55 / 91.88
License year	93.06 / 90.66	96.78 / 97.29	96.85 / 98.21	95.73 / 94.53	93.56 / 89.95
State	94.24 / 90.58	96.72 / 97.45	98.06 / 98.27	96.06 / 95.08	94.50 / 92.21
City	88.98 / 85.68	95.25 / 96.14	96.58 / 96.95	91.76 / 91.00	91.65 / 89.69
County	83.36 / 80.28	95.14 / 89.78	96.65 / 96.97	86.31 / 85.61	87.31 / 85.42

Agreement with the cleaned validation sample; scan-verified accuracy not assessed. Same frames, links and scorer as Table 2. Name is a link key, so its agreement is linkage-conditioned. Years as two-digit tokens; school codes after syntactic normalization; places after formatting normalization. Integer counts for 1921 appear in Table G.2.

Table G.2 gives the integer counts behind Table G.1: agreeing linked values, scorable output values and scorable reference values, for each field and the four workflows shown on the 1921 analysis frame. Table G.4 gives the disagreement counts of Section 5.3. Table G.3 shows what release cleaning does to the release table's agreement, scored as read and after cleaning on links fixed from the as-read values. We selected the cleaning rules by comparing candidates against the transcription in the 1912 and 1921 editions before running this comparison.

Table G.2. Field counts by workflow, 1921: agreeing / scorable output / scorable reference values

Field	MLLM	Parser	Combined	Open LLM
Name	116,883 / 122,953 / 123,708	120,998 / 125,101 / 123,708	122,032 / 124,477 / 123,708	118,070 / 122,915 / 123,708
Birth year	105,239 / 110,774 / 109,480	106,823 / 109,749 / 109,480	107,624 / 109,825 / 109,480	104,037 / 108,043 / 109,480
Medical school	108,682 / 114,831 / 121,137	117,191 / 120,890 / 121,137	118,746 / 121,851 / 121,137	112,038 / 118,578 / 121,137
Graduation year	112,852 / 120,019 / 118,984	115,252 / 120,764 / 118,984	116,894 / 121,920 / 118,984	112,374 / 117,178 / 118,984
License year	93,603 / 100,580 / 103,249	100,454 / 103,791 / 103,249	101,401 / 104,701 / 103,249	97,604 / 101,961 / 103,249
State	112,472 / 119,344 / 124,165	120,998 / 125,101 / 124,165	122,013 / 124,424 / 124,165	118,053 / 122,897 / 124,165
City	106,196 / 119,344 / 123,948	119,158 / 125,101 / 123,948	120,164 / 124,424 / 123,948	112,788 / 122,915 / 123,948
County	99,482 / 119,344 / 123,925	111,256 / 116,944 / 123,925	120,170 / 124,334 / 123,925	106,087 / 122,915 / 123,925

Agreement with the cleaned validation sample; scan-verified accuracy not assessed. Unmatched rows' values stay in the denominators; a cell with several years or schools contributes each value. Source: the field-agreement comparison.

Table G.3. Release cleaning and agreement, combined table, 1921 (precision / recall, %)

Field	As read	After cleaning
Medical school	97.45 / 98.03	97.48 / 98.11
Birth year	98.00 / 98.30	98.01 / 98.30
City	96.58 / 96.95	96.58 / 96.95
County	90.90 / 90.96	96.65 / 96.97

Agreement with the cleaned validation sample; scan-verified accuracy not assessed. Links fixed from the as-read values. The as-read school, birth-year and city figures and the cleaned county figure are the combined table's entries in Table G.1; the release table's `*_as_read` columns against its cleaned columns. Source: the release-cleaning comparison.

Table G.4. Disagreement between the MLLM workflow and the AI-assisted parser on rows linked through the combined table, 1921

Field	Both agree	Only MLLM agrees	Only parser agrees	Neither agrees
Birth year	105,353	381	1,893	64
Graduation year	112,659	1,228	3,031	118
License year	95,196	534	5,676	120
Medical school	108,471	1,140	7,106	318

Agreement with the cleaned validation sample; scan-verified accuracy not assessed. Cells on the 122,032 linked rows of the combined table on 1,034 pages; an absent extraction entry counts as non-agreement. Counts describe agreement with the transcription and do not establish which extraction reproduces the printed entry. Rows where the reference is blank are left out; a cell with several school codes counts once here and once per value in Table G.2. Source: the comparison of extraction disagreements.

H Cost accounting

Table H.1 gives the cost of manual transcription and the direct processing charges for the machine workflows, in total and per page. Manual costs are the delivered workbook page counts, 2,242 pages across both editions, at the invoiced €14.70 per page. The machine cost per page divides the combined Gemini and Azure charges by the same page count. We allocate Gemini charges by whole billing days and assume one successful Azure OCR pass per page, so the machine figures are estimates, not invoice totals. Ratios use unrounded inputs and one exchange-rate date (ECB, 3 February 2026).

The comparison covers direct processing charges only. It leaves out engineering, review, model development, coding-agent assistance, energy and the local computing hardware. The hardware is shared with other project work and hosted-API workflows do not need it, so adding it would not make the comparison like for like. The manual transcription also covers a different set of pages and fields than the machine outputs, and we do not estimate a separate cost for the open model pipelines.

Table H.1. Manual transcription and estimated machine processing charges

Component	Unit	1912	1921	Both editions
Purchased transcription	Pages	1,007	1,235	2,242
Manual transcription	EUR	14,802.90	18,154.50	32,957.40
Gemini, allocated billing	DKK	136.99	241.71	378.70
Azure, one-pass estimate	USD	13.88	14.80	28.68
Gemini plus Azure	EUR	30.10	44.90	75.01
Manual cost per page	EUR	14.70	14.70	14.70
Machine cost per page	EUR	0.03	0.04	0.03
Manual / estimated API cost	Ratio	492	404	439

Manual costs use delivered workbook page counts at €14.70 per page, reconciled to PDS invoices; the 1918 batch and unassigned supplement are excluded. These are purchased-coverage counts, not the analysis frames. API figures cover Gemini and Azure, not a separate cost estimate for the open model pipelines.

Gemini 2.0 Flash billing is allocated by whole days: 28–31 January 2026 (1912), 23–25 January and 3–5 February (1921); other-model charges are excluded. Edition overlap on 25 and 31 January makes attribution approximate. Excluding 31 January gives DKK 114.12 for 1912; excluding 25 January gives DKK 185.45 for 1921. These are scenarios, not confidence bounds.

Azure assumes 1,388 and 1,480 successful one-page images at the conditional West Europe prebuilt-layout S0 list price of USD 10 per 1,000 pages (effective from August 2021). Actual Azure billing and repeat-call counts were not available; these are list-price estimates, not invoice totals. EUR equivalents and ratios use ECB reference rates on 3 February 2026: EUR1 = DKK 7.4687 = USD 1.1801, a common-date presentation convention; totals and ratios use unrounded inputs.

I Robustness to the page frame

The analysis frames keep every transcribed page without an ignored-row placeholder (Section 4.1). Two choices could move the results: leaving out the pages flagged for incomplete coverage, and keeping the pages seen during parser development, drawn into the frozen design sample or used in diagnostics. Table I.1 scores every workflow on all transcribed pages and on the incomplete pages alone; Table I.2 scores them on the analysis frame without the exposed pages and on the exposed pages alone. Links are fixed edition-wide before any page restriction, so each frame scores the same links.

Table I.1. Record agreement by coverage frame, 1912 and 1921

Coverage	1912				1921			
	All transcribed		Incomplete pages		All transcribed		Incomplete pages	
Workflow	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)
MLLM	94.76	94.21	93.61	92.40	92.55	93.96	84.11	92.90
Parser	97.48	97.27	96.89	96.48	95.11	97.44	86.54	97.42
Combined	97.65	97.72	97.03	96.93	96.43	98.32	87.89	98.53
Open LLM	97.57	95.60	96.14	93.15	94.44	95.22	85.90	96.00
Open OCR-LLM	96.42	93.54	95.77	93.21	93.35	92.74	85.02	92.85

Agreement with the cleaned validation sample; scan-verified accuracy not assessed. All transcribed: 1,233 pages and 145,064 reference rows in 1921, 1,006 and 137,501 in 1912. Incomplete pages: the pages with an ignored-row placeholder, 197 pages and 20,897 rows in 1921, 103 and 13,775 in 1912. On these pages part of the printed entries were not transcribed, so precision understates agreement. Links after the second pass; open LLM and open OCR-LLM as in Table 2.

Table I.2. Record agreement with and without exposed pages, 1912 and 1921

Exposure	1912				1921			
	Without exposed pages		Exposed pages only		Without exposed pages		Exposed pages only	
Workflow	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)
MLLM	95.01	94.64	94.05	92.83	94.33	94.29	92.74	93.12
Parser	97.62	97.46	96.96	96.66	96.59	97.46	97.59	97.39
Combined	97.80	97.89	97.20	97.20	98.06	98.31	97.87	98.10
Open LLM	97.78	95.94	97.36	95.31	96.09	95.32	95.84	93.57
Open OCR-LLM	96.55	93.64	96.07	93.16	94.87	92.67	95.28	93.10

Agreement with the cleaned validation sample; scan-verified accuracy not assessed. Without exposed pages: the analysis frame less every page with a scan seen while tuning the parser, drawn into the frozen design sample or used in diagnostics, 902 pages and 107,565 reference rows in 1921, 794 and 108,531 in 1912. Exposed pages only: the 134 and 109 pages left out there (16,602 and 15,195 rows): in 1921, 16 design, 116 tuning and 2 diagnostic pages; in 1912, 16, 93 and 0. A page on several lists counts once, as design before tuning before diagnostic. The record of pages seen during parser development is incomplete for 1921. Links after the second pass.