Computational Humanities Research

www.cambridge.org/chr

Retrieving information from unstructured historical sources using large language models

Spencer Dean Stewart[1] and Sanskriti Sinha[2]

[1]Libraries and School of Information Studies, Purdue University, West Lafayette, USA and [2]Department of Statistics, Purdue University, West Lafayette, USA

Short Article

Cite this article: Stewart Spencer Dean and Sanskriti Sinha. 2025. "Retrieving information from unstructured historical sources using large language models" *Computational Humanities Research*, **1**:e17, https://doi.org/10.1017/chr.2025.10019

Received: 31 January 2025
Revised: 14 October 2025
Accepted: 4 November 2025

Keywords:
cultural analytics; digital humanities; historical text processing; information extraction; large language models

Corresponding authors:
Spencer Dean Stewart;
Email: stewa443@purdue.edu

Abstract

The volumes of historical data locked behind unstructured formats have long been a challenge for researchers in the computational humanities. While optical character recognition (OCR) and natural language processing have enabled large-scale text mining projects, the irregular formatting, inconsistent terminology and evolving printing practices complicate automated parsing and information extraction efforts for historical documents. This study explores the potential of large language models (LLMs) in processing and structuring irregular and non-standardized historical materials, using the U.S. Department of Agriculture's Plant Inventory books (1898–2008) as a test case. Given the frequent evolution of these historical records, we implemented a pipeline combining OCR, custom segmentation rules and LLMs to extract structured data from the scanned texts. It provides an example of how incorporating LLMs into data-processing pipelines can enhance the accessibility and usability of historical and archival materials for scholars.

Plain language summary

Historical documents often contain valuable information, but their unstructured and inconsistent formats make them difficult to use for large-scale analysis. Traditional approaches to information extraction, such as regular expressions and rule-based natural language processing tools, often struggle with these irregularities. This research explores how large language models (LLMs) can help scholars extract and organize data from challenging historical materials. We use the U.S. Department of Agriculture's Plant Inventory books (1898–2008) as a case study. These volumes feature varied layouts, old fonts and inconsistent terminology.

Our approach involved a three-step process that combined the strengths of traditional and newer extraction methods:

1. **Optical character recognition:** We used Azure AI Document Intelligence to convert scanned book images into raw text.
2. **Custom segmentation:** We developed a specialized script to divide this text into logical blocks to handle entries that spanned multiple columns or pages.
3. **LLM testing:** We evaluated commercial and open-source LLMs on their ability to extract specific information (e.g., plant names, type and origins) from the segmented text and structure it into an organized JSON format.

We found that performance varied across models. Open-source models, such as Llama, were constrained by context window limits when processing longer passages. While commercial LLMs showed improved accuracy in identifying and extracting relevant information, the newest models did not necessarily yield the best results. Overall, our findings suggest that LLMs provide a promising starting point for making unstructured historical data more accessible to computational humanities research.

Introduction

Historical sources contain vast amounts of unstructured data with immense potential for scholarly analysis, yet their inherent complexities make it difficult to efficiently capture and process relevant information. Beyond the initial challenge of converting analog materials into machine-readable formats, variability in language, structure and context further complicates efforts to extract insights – even when using state-of-the-art named entity recognition (NER) tools (Ehrmann et al. 2023). In the past, building large-scale historical databases has often relied on the labor of countless workers who painstakingly compiled, digitized and standardized heterogeneous records. Because such projects require extensive human intervention for processing and annotation, many promising initiatives never materialize due to the lack of sufficient resources. As a result, significant portions of historical texts remain underutilized and locked behind barriers of unstructured complexity.

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Large language models (LLMs) offer promising solutions to these challenges, assisting in tasks ranging from data extraction and annotation to contextual analysis and summarization. A growing number of studies have explored using LLMs for text classification and data mining, with significant implications for fields like healthcare (Fink et al. 2023; Moulaei et al. 2024), the social sciences (Abdurahman et al. 2025; Ziems et al. 2024) and the humanities (Laato et al. 2025; Ma et al. 2024). Targeted approaches for capturing specific entities or classifying texts have proven effective (Bamman et al. 2024; Polak and Morgan 2024; Underwood 2025), and some studies suggest that ChatGPT can even outperform crowd-sourced workers on some annotation tasks (Gilardi, Alizadeh, and Kubli 2023; Törnberg 2023). Recently, Stuhler, Ton, and Ollion (2025) argued that generative AI tools allow for a more flexible approach to data mining by turning deterministic codebooks into adaptable "promptbooks."

We build on this body of research to evaluate the potential of LLMs for targeted text mining and information extraction, focusing on the creation of a database from the *Plant Inventory* booklets published by the U.S. Department of Agriculture (USDA) between 1898 and 2008. The unstructured and inconsistent formatting of the records poses challenges for traditional information extraction methods, such as regular expressions, rule-based systems and traditional machine learning-based parsers. The variable formatting, however, makes them an ideal test case for assessing the reliability of LLMs for mining unstructured historical texts for humanities research. We find that prompt-based approaches to data mining provide a more accessible and flexible approach to extracting information from unstructured historical texts. However, the "black-boxed" nature of LLMs can lead to errors that are both non-random and ambiguous, especially when working with commercial models and historical texts (Stuhler, Ton, and Ollion 2025; Underwood, Nelson, and Wilkens 2025; Zhang et al. 2024). This tradeoff requires a greater degree of care and consideration when validating the reliability of information extraction with generative AI models, especially in fields like digital humanities that are increasingly dedicated to transparency and reproducibility (Ries, van Dalen-Oskam, and Offert 2024).

Data

Our primary source of data is the *Plant Inventory* series, published from 1898 to 2008 by the USDA documenting the introduction of over 650,000 plant materials into the U.S. National Plant Germplasm System. While there was a longer tradition of introducing foreign plants into the United States, the establishment of the USDA's "Section of Seed and Plant Introduction" in 1898 marked a turning point in centralizing and expanding this process. While the division underwent multiple changes over the years, including its name, it became an important organ of promoting the introduction of foreign germplasm by agricultural explorers, such as David Fairchild and Niels Hansen. The impact of these efforts can be seen throughout the American agricultural landscape, including the countless fields of hard winter wheat, soybeans and lemons. To systematically document newly imported plants and make them available to the growing number of experiment stations scattered throughout the United States, the USDA began publishing this information in booklets that would come to be known as *Plant Inventory*. Within these booklets, each plant was given a unique number to serve as a means of designation due to the complicated and multilingual context of global agriculture. Each entry also listed detailed information about the plants, including botanical and common names, collection location, collector's name and descriptive notes about its potential uses or climatic suitability. This systematic documentation reflected the USDA's effort to centralize agricultural experimentation and maintain a robust archive for plant genetic resources (Fullilove 2019; Hodge and Erlanson 1956; Hyland 1977; Li 2022).

From 1898 to 2008, over 200 booklets were published as bound paper volumes. Initial publications were irregular, but they appeared quarterly from 1903 to 1941, shifting to an annual schedule in 1942. Beginning in 2009, they converted to a digital model. As the regularity of the publications shifted, so too did the format and information included. As seen in Figure 1, the differences could be dramatic, including the introduction of columns on the page or lumping together a range of numbers that might "continue" onto subsequent columns or pages. Such changes introduce additional complexities when trying to digitize the materials. These booklets, then, provide not only a window into the global origins of U.S. agriculture but also a case study for testing the limits of existing optical character recognition (OCR) tools and LLMs in processing complex historical documents.

Workflow

Given the complexity of our documents, which feature multiple formatting styles and irregular text structures, we developed a multi-stage workflow that includes the steps represented in Figure 2. The process begins by converting the PDF image into raw OCR output. This unstructured text is then embedded within a specific LLM prompt, which instructs the model to return the information as a structured final JSON output. A more detailed overview of this process is as follows.

Optical character recognition

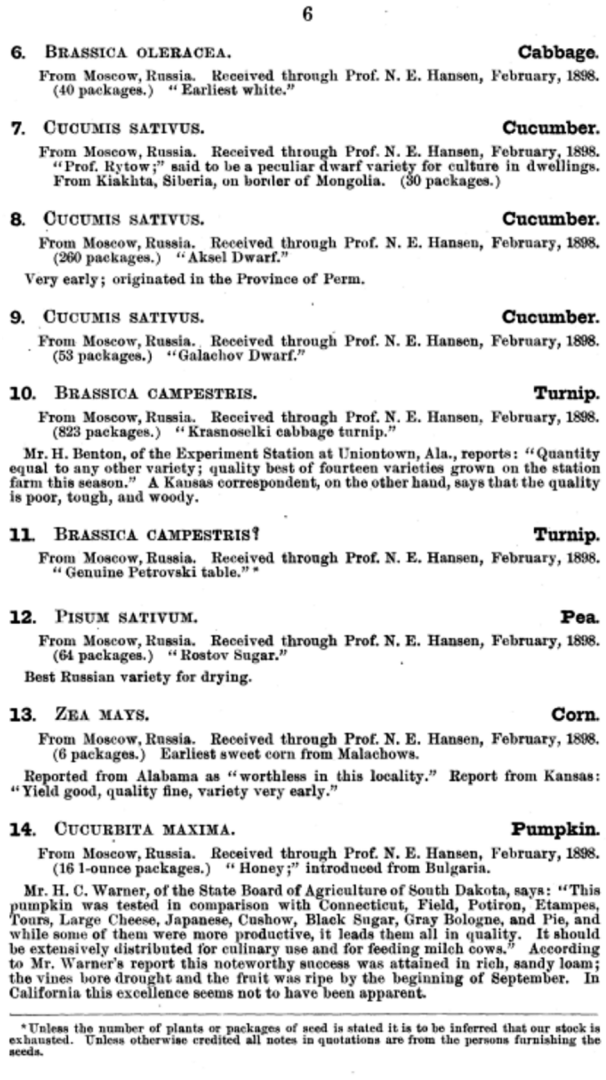
The Plant Inventory books have already been scanned and are available online through multiple venues, including the USDA Agricultural Research Service and the National Agricultural Library. We downloaded a complete run of the 231 publications through the Agricultural Research Service, which totaled 45,051 pages (https://www.ars.usda.gov/northeast-area/beltsville-md-barc/beltsville-agricultural-research-center/national-germplasm-resources-laboratory/docs/plant-inventory-books/).

We experimented with multiple OCR solutions, each presenting unique strengths and limitations for historical documents. We ultimately settled on using Azure AI Document Intelligence, Microsoft's OCR engine, after experimenting with both Tesseract and Google Vision OCR. We found that while Tesseract performed adequately on clean images, its handling of complex layouts, such as multi-column pages, resulted in omissions or fragmented text. Google Vision OCR demonstrated robust capabilities that were comparable to Azure, but its ecosystem was less conducive to rapid, iterative prototyping. In our side-by-side comparison during initial testing, Azure consistently provided coherent and complete text extractions with minimal errors, including when dealing with irregularities, such as inconsistent spacing and partial-page text alignments. Its multi-column detection capabilities significantly minimized the need for extensive manual intervention in preparing the raw OCR output for subsequent processing.

LLM prompting

preparing the text for prompting first required us to segment the text into manageable blocks. Simple regular expressions proved

https://doi.org/10.1017/chr.2025.10019 Published online by Cambridge University Press



6

6. BRASSICA OLERACEA. Cabbage.

From Moscow, Russia. Received through Prof. N. E. Hansen, February, 1898. (40 packages.) "Earliest white."

7. CUCUMIS SATIVUS. Cucumber.

From Moscow, Russia. Received through Prof. N. E. Hansen, February, 1898. "Prof. Rytow;" said to be a peculiar dwarf variety for culture in dwellings. From Kiakhta, Siberia, on border of Mongolia. (30 packages.)

8. CUCUMIS SATIVUS. Cucumber.

From Moscow, Russia. Received through Prof. N. E. Hansen, February, 1898. (260 packages.) "Aksel Dwarf."

Very early; originated in the Province of Perm.

9. CUCUMIS SATIVUS. Cucumber.

From Moscow, Russia. Received through Prof. N. E. Hansen, February, 1898. (53 packages.) "Galachov Dwarf."

10. BRASSICA CAMPESTRIS. Turnip.

From Moscow, Russia. Received through Prof. N. E. Hansen, February, 1898. (823 packages.) "Krasnoselki cabbage turnip."

Mr. H. Benton, of the Experiment Station at Uniontown, Ala., reports: "Quantity equal to any other variety; quality best of fourteen varieties grown on the station farm this season." A Kansas correspondent, on the other hand, says that the quality is poor, tough, and woody.

11. BRASSICA CAMPESTRIS? Turnip.

From Moscow, Russia. Received through Prof. N. E. Hansen, February, 1898. "Genuine Petrovski table." *

12. PISUM SATIVUM. Pea.

From Moscow, Russia. Received through Prof. N. E. Hansen, February, 1898. (64 packages.) "Rostov Sugar."

Best Russian variety for drying.

13. ZEA MAYS. Corn.

From Moscow, Russia. Received through Prof. N. E. Hansen, February, 1898. (6 packages.) Earliest sweet corn from Malachows.

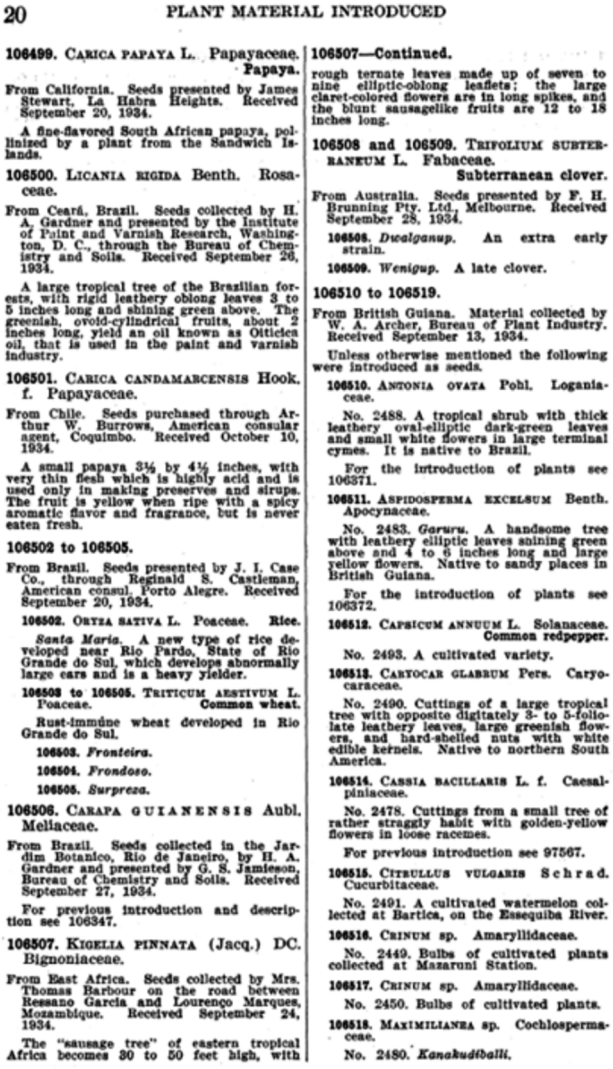
Reported from Alabama as "worthless in this locality." Report from Kansas: "Yield good, quality fine, variety very early."

14. CUCURBITA MAXIMA. Pumpkin.

From Moscow, Russia. Received through Prof. N. E. Hansen, February, 1898. (16 1-ounce packages.) "Honey;" introduced from Bulgaria.

Mr. H. C. Warner, of the State Board of Agriculture of South Dakota, says: "This pumpkin was tested in comparison with Connecticut, Field, Potiron, Etampes, Tours, Large Cheese, Japanese, Cushow, Black Sugar, Gray Bologne, and Pie, and while some of them were more productive, it leads them all in quality. It should be extensively distributed for culinary use and for feeding milch cows." According to Mr. Warner's report this noteworthy success was attained in rich, sandy loam; the vines bore drought and the fruit was ripe by the beginning of September. In California this excellence seems not to have been apparent.

*Unless the number of plants or packages of seed is stated it is to be inferred that our stock is exhausted. Unless otherwise credited all notes in quotations are from the persons furnishing the seeds.



20 PLANT MATERIAL INTRODUCED

106499. CARICA PAPAYA L. Papayaceae. Papaya.

From California. Seeds presented by James Stewart, La Habra Heights. Received September 20, 1934.

A fine-flavored South African papaya, pollinized by a plant from the Sandwich Islands.

106500. LICANIA RIGIDA Benth. Rosaceae.

From Ceará, Brazil. Seeds collected by H. A. Gardner and presented by the Institute of Paint and Varnish Research, Washington, D. C., through the Bureau of Chemistry and Soils. Received September 26, 1934.

A large tropical tree of the Brazilian forests, with rigid leathery oblong leaves 3 to 5 inches long and shining green above. The greenish, ovoid-cylindrical fruits, about 2 inches long, yield an oil known as Oiticica oil, that is used in the paint and varnish industry.

106501. CARICA CANDAMARCENSIS Hook. f. Papayaceae.

From Chile. Seeds purchased through Arthur W. Burrows, American consular agent, Coquimbo. Received October 10, 1934.

A small papaya 3½ by 4½ inches, with very thin flesh which is highly acid and is used only in making preserves and sirups. The fruit is yellow when ripe with a spicy aromatic flavor and fragrance, but is never eaten fresh.

106502 to 106505.

From Brazil. Seeds presented by J. I. Case Co., through Reginald S. Castleman, American consul, Porto Alegre. Received September 20, 1934.

106502. ORYZA SATIVA L. Poaceae. Rice.

Santa Maria. A new type of rice developed near Rio Pardo, State of Rio Grande do Sul, which develops abnormally large ears and is a heavy yielder.

106503 to 106505. TRITICUM AESTIVUM L. Poaceae. Common wheat.

Rust-immune wheat developed in Rio Grande do Sul.

106503. *Fronteira.*

106504. *Frondoso.*

106505. *Surpreza.*

106506. CARAPA GUIANENSIS Aubl. Meliaceae.

From Brazil. Seeds collected in the Jardim Botanico, Rio de Janeiro, by H. A. Gardner and presented by G. S. Jamieson, Bureau of Chemistry and Soils. Received September 27, 1934.

For previous introduction and description see 106347.

106507. KIGELIA PINNATA (Jacq.) DC. Bignoniaceae.

From East Africa. Seeds collected by Mrs. Thomas Barbour on the road between Ressano Garcia and Lourenço Marques, Mozambique. Received September 24, 1934.

The "sausage tree" of eastern tropical Africa becomes 30 to 50 feet high, with

106507—Continued.

rough ternate leaves made up of seven to nine elliptic-oblong leaflets; the large claret-colored flowers are in long spikes, and the blunt sausagelike fruits are 12 to 18 inches long.

106508 and 106509. TRIFOLIUM SUBTERRANEUM L. Fabaceae. Subterranean clover.

From Australia. Seeds presented by F. H. Brunning Pty. Ltd., Melbourne. Received September 28, 1934.

106508. *Dwalganup.* An extra early strain.

106509. *Wenigup.* A late clover.

106510 to 106519.

From British Guiana. Material collected by W. A. Archer, Bureau of Plant Industry. Received September 13, 1934.

Unless otherwise mentioned the following were introduced as seeds.

106510. ANTONIA OVATA Pohl. Loganiaceae.

No. 2488. A tropical shrub with thick leathery oval-elliptic dark-green leaves and small white flowers in large terminal cymes. It is native to Brazil.

For the introduction of plants see 106371.

106511. ASPIDOSPERMA EXCELSUM Benth. Apocynaceae.

No. 2483. *Garuru.* A handsome tree with leathery elliptic leaves shining green above and 4 to 6 inches long and large yellow flowers. Native to sandy places in British Guiana.

For the introduction of plants see 106372.

106512. CAPSICUM ANNUUM L. Solanaceae. Common redpepper.

No. 2493. A cultivated variety.

106513. CARYOCAR GLABRUM Pers. Caryocaraceae.

No. 2490. Cuttings of a large tropical tree with opposite digitately 3- to 5-foliolate leathery leaves, large greenish flowers, and hard-shelled nuts with white edible kernels. Native to northern South America.

106514. CASSIA BACILLARIS L. f. Caesalpiniaceae.

No. 2478. Cuttings from a small tree of rather straggly habit with golden-yellow flowers in loose racemes.

For previous introduction see 97567.

106515. CITRULLUS VULGARIS Schrad. Cucurbitaceae.

No. 2491. A cultivated watermelon collected at Bartica, on the Essequibo River.

106516. CRINUM sp. Amaryllidaceae.

No. 2449. Bulbs of cultivated plants collected at Mazaruni Station.

106517. CRINUM sp. Amaryllidaceae.

No. 2450. Bulbs of cultivated plants.

106518. MAXIMILIANEA sp. Cochlospermaceae.

No. 2480. *Kanakudiballi.*

Figure 1. Page examples from Plant Inventory booklets.

insufficient for this task due to the complexity and inconsistency of the entries, which often lacked clear delimiters and sometimes spanned multiple PI numbers; a new regular expression configuration would have been required for nearly every printed volume. Additionally, while sending one OCR-processed page at a time avoided overloading the LLM and minimized hallucination, it introduced a new challenge: many entries extended across multiple pages, making single-page segmentation inadequate (for some examples of these complexities, see Appendix A).

To address this, we developed a Python script to identify and reconstruct multi-page entries by analyzing textual and formatting patterns across all 231 booklets. The script began by detecting anchor elements, such as bold PI numbers or numeric ranges (e.g., "300968–300972") that typically marked the start of a new entry. It then scanned for continuation cues, such as the keywords "continued," "to," or conjunctions like "and," all of which often signaled entries that spanned multiple PI numbers or pages.

This process distinguished between true continuations (e.g., "300968–300972 – Continued") and independent entries with similar structures. It also handled complex edge cases, including nested continuations and synonyms used inconsistently to indicate entry extension. To merge these segments, the script maintained a dynamic buffer: one set of functions scanned each line for trigger terms, while another group handled the logic of whether to append content to an ongoing record or start a new one based on contextual cues. After iterative testing and refinement, we established a set of heuristics capable of reliably merging fragmented entries.

Once the text was segmented into coherent multi-page blocks, we used OpenAI's API playground to craft a prompt that would process the text in accordance with our criteria. The main data points that we extracted included:

- **PI** – The unique plant identifier number assigned to the crop.
- **Name** – The scientific name of the crop species.
- **Type** – The botanical family or classification of the crop.
- **Origin** – The geographic location where the crop specimen was collected or reported.
- **Date** – The date the specimen was received or entered into the record.
- **Presented** – The individual or institution who submitted or presented the crop.

https://doi.org/10.1017/chr.2025.10019 Published online by Cambridge University Press



Figure 2. Pipeline for structured data extraction using OCR and LLM prompting.

Evaluation

To evaluate the prompt's extraction performance, we created a ground-truth dataset by manually labeling 525 crop entries from 27 sample text chunks. Our evaluation criteria were designed to provide a granular analysis of performance, moving beyond a single similarity score that might obscure important distinctions between minor transcription errors and complete extraction failures. We therefore employed the following metrics:

1. **Output reliability:** We measured the percentage of model outputs that could be successfully parsed as valid JSON. This metric assessed a model's ability to consistently adhere to the required output format.
2. **Field-level accuracy:** We calculated accuracy for each metadata field (e.g., *name* and *type*) by comparing the extracted value against the ground truth. For this, we first normalized both sets of text to account for minor differences. This normalization included converting text to lowercase, removing punctuation (e.g., "CUCUMIS MELO" vs. "Cucumis melo."), and reconciling over- or under-extraction of entities (e.g., "Col. Robert H. Montgomery…" vs. "Robert H. Montgomery").
3. **Computational efficiency:** We measured the average tokens consumed and the runtime per file for each model. These metrics provide insight into the computational cost and practicality of each approach alongside its extraction accuracy.

We evaluated a suite of state-of-the-art language models available in late 2024 on their ability to extract structured crop metadata from OCR-transcribed documents. Following an initial round of peer review, we updated our analysis to include several more recent models available by mid-2025. Our comparison included models from OpenAI (GPT-5, GPT-4o, GPT-4o-mini and GPT-3.5), Google (Gemini 2.5 Flash and Pro, and Gemini 1.5 Flash and Pro) and Meta (Llama3-8B), with the temperature set to 0 (when available) to ensure more deterministic outputs. While most models were accessed via commercial APIs, the Llama3-8B model was run locally on a Google Colab A100 High-RAM GPU instance. This local setup introduced additional overhead in managing dependencies and runtime stability, and both this and the Gemini 2.5 Pro failed to successfully process the full set of 27 test files.

An overview of the results, as shown in Table 1, reveals a clear tradeoff between model accuracy, speed and token efficiency. GPT-4o emerged as the top-performing model, achieving the highest accuracy in the fields of Name (98.66%) and Origin (97.13%). The Gemini 1.5 and 2.5 series models were highly competitive, though they showed variability in certain fields; for instance, Gemini 2.5 Flash scored lower on Presenter accuracy (74.19%). In terms of processing speed, GPT-3.5 Turbo was the most efficient model at 15.20 seconds per file, with Gemini 1.5 Flash following at 20.03 seconds.

A qualitative review of the discrepancies reveals several recurring error patterns that highlight the specific challenges of this extraction task. For instance, in the *type* field, models frequently struggled with semantic ambiguity; for a single entry, one model might extract a generic common name like "tree," while another would extract the scientific family name, both of which could be technically correct but would fail a direct comparison against a gold standard expecting "acacia." A second common issue was the over-extraction of structured metadata, particularly in the *presenter* field, where models would correctly identify a name but also include their title and affiliation (e.g., "Col. Robert H. Montgomery, Director, Coconut Grove arboretum"). Finally, we

https://doi.org/10.1017/chr.2025.10019 Published online by Cambridge University Press

Table 1. Comparison of large language models on extraction accuracy, speed and cost

Model	Success (n = 27)	PI (%)	Name (%)	Type (%)	Origin (%)	Date (%)	Presenter (%)	Avg. Tokens	Avg. Time (s)
gemini-1.5-flash-002	27	100.00	95.01	70.06	87.72	100.00	94.82	4,466.85	20.03
gemini-1.5-pro-002	27	100.00	97.51	87.19	92.54	92.93	86.62	4,535.59	29.43
gemini-2.5-flash	27	100.00	97.13	86.81	89.10	98.85	74.19	12,961.67	59.17
gemini-2.5-pro	23	100.00	96.48	84.36	89.87	98.68	81.72	10,492.74	61.67
gpt-3.5-turbo	27	100.00	89.59	68.10	65.84	59.05	60.86	3,789.00	15.20
gpt-4o	27	100.00	98.66	83.72	97.13	100.00	94.64	4,260.63	27.83
gpt-4o-mini	27	100.00	94.33	81.72	83.19	78.78	79.20	4,078.33	48.64
gpt-5-2025-08-07	27	100.00	95.57	87.28	88.25	98.46	83.82	10,067.85	95.19
gpt-5-mini-2025-08-07	27	100.00	96.48	85.32	90.02	93.74	84.54	8,323.30	98.87
gpt-5-nano-2025-08-07	27	100.00	95.17	76.68	87.82	86.55	84.24	16,050.41	82.59
llama3-8b	17	100.00	96.15	83.65	78.37	77.88	79.33	3,512.48	67.15

Note: "Success" refers to the number of files that the model was able to successfully process and return in the required JSON format. Subsequent results are based solely on the successfully-returned files. Overall accuracy is calculated by comparing the output from each model to a "gold standard" sample after standardizing for casing, punctuation and over-extraction of metadata.

observed frequent cases of embedded descriptive phrases, especially in the *origin* field, where text like "at 12,000 feet altitude" was inserted into the middle of a place name. These examples demonstrate that failures were often not simple factual mistakes but rather issues of formatting, specificity and interpretation. While our post-processing framework was ultimately able to manage a majority of these errors, further prompt refinement and model fine-tuning would likely improve the consistency of the initial extraction (Brown et al. 2020).

Discussion

One key consideration in our process was the tradeoff between commercial and open-source language models. Commercial APIs (e.g., OpenAI and Gemini) typically offer ease of use, powerful performance, larger context windows and scalable infrastructure. However, they introduce concerns around consistency and reproducibility. Calling the same prompt twice may return different outputs, and proprietary models provide limited visibility into versioning, training data or system behavior. These challenges make it difficult to ensure transparency and replicability.

For this reason, many researchers in the humanities and social sciences have turned to open-source alternatives, such as Meta's LLaMA models (Alizadeh et al. 2025; Ziems et al. 2024). These models offer greater transparency and control, but they also demand significantly more computational power. As early-career researchers in Digital Humanities, our access to such resources was limited. We were able to run the LLaMA 3 8B model with a monthly subscription to Google Colab Pro, which was sufficient for small-scale evaluations. However, attempts to scale up to larger models like LLaMA 70B quickly ran into memory limitations, constraints that commercial API-based models typically circumvent. For researchers without access to institutional computing clusters, Colab Pro remains a practical option for working with smaller open models. Still, as of the writing of this paper, these models have their own drawbacks, including inconsistent performance on complex tasks and smaller context windows compared to commercial alternatives.

Moreover, our study provides a limited but productive glimpse into the trajectory of model change over time, along with how prompts can be model sensitive. When holding our prompt constant, the generational leap from older models like GPT-3.5 Turbo to the more recent GPT-4o is unambiguous as seen in Table 1, with accuracy scores jumping more than 30 percentage points in fields like *origin*. However, the progression among newer models is less linear. The performance of the "GPT-5" series, for instance, did not uniformly surpass that of GPT-4o on this task and came with a substantial increase in processing time. Factors, such as model size (mini vs. standard), intended function (flash vs. pro) and the incomplete runs of Llama3-8B and Gemini 2.5 Pro, are all important to consider when identifying models for information extraction. While this might point to a general yet uneven increase in capacity over time, it also underscores the challenge of prompt portability: a prompt optimized for GPT-4o will not necessarily yield promising results with Gemini 2.5 Pro.

Ultimately, LLMs are not a magic bullet for data extraction. They work best when paired with traditional methods, such as manual annotation, human-in-the-loop verification and domain-specific heuristics (Khraisha et al. 2024; Schilling-Wilhelmi et al. 2025; Schroeder, Jaldi, and Zhang 2025). By designing thoughtful validation strategies (e.g., checking field alignment, comparing PI numbers across pages and calculating similarity scores), we can build confidence in the outputs, even when working with the black-box nature of generative AI. Used strategically, LLMs can accelerate labor-intensive processes like historical metadata extraction while still preserving data integrity.

Conclusion

This study demonstrates the promise and limitations of using LLMs to extract structured metadata from unstructured historical documents. By combining OCR, custom segmentation heuristics and LLM prompting, we developed a pipeline capable of transforming over 45,000 pages of USDA *Plant Inventory* records into a structured and analyzable dataset. Our multi-model evaluation showed that commercial models like OpenAI and Gemini offer high accuracy and ease of deployment, while open-source models like LLaMA provide transparency and control, though at greater computational cost.

https://doi.org/10.1017/chr.2025.10019 Published online by Cambridge University Press

Our results also highlight the indispensable role of validation and post-processing in any LLM-based workflow. Even the most powerful and costly models produced formatting glitches, hallucinated entries and occasional edge-case failures. LLMs, therefore, are complements – not replacements – for human expertise: domain specialists will continue to be essential for designing prompts, interpreting ambiguous outputs and enforcing quality standards. When generative AI is paired with systematic checks and subject knowledge, it has the potential to deliver more reliable and scalable results.

Data availability statement. All code, data and examples can be found on our GitHub repository here: https://github.com/sdstewart/Plant-Inventory-LLMs/.

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the United States.

Author contributions. Conceptualization: S.S., S.D.S.; Data curation: S.S.; Methodology: S.S., S.D.S.; Writing – original draft: S.S., S.D.S.; Writing – review and editing: S.S., S.D.S. Both authors approved the final submitted draft.

Funding statement. This research was supported by the Libraries and School of Information Studies at Purdue University.

Competing interests. The authors declare none.

Disclosure of use of AI tools. Generative AI was used for assistance in prompt refinement, generating and debugging code, and in basic formatting. Otherwise, no AI tools were used outside of the experimental methodology outlined in the paper.

References

Abdurahman, Suhaib, Alireza Salkhordeh Ziabari, Alexander K. Moore, Daniel M. Bartels, and Morteza Dehghani. 2025. "A Primer for Evaluating Large Language Models in Social-Science Research." *Advances in Methods and Practices in Psychological Science* 8, no. 2: 25152459251325174.

Alizadeh, Meysam, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan D. Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2025. "Open-Source LLMs for Text Annotation: A Practical Guide for Model Setting and Fine-Tuning." *Journal of Computational Social Science* 8, no. 1: 17.

Bamman, David, Kent K. Chang, Li Lucy, and Naitian Zhou. 2024. "On Classification with Large Language Models in Cultural Analytics." In *Proceedings of the Computational Humanities Research Conference 2024,* edited by Wouter Haverals, Marijin Koolen, and Laure Thompson, 494–527. CEUR Workshop Proceedings.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei 2020. "Language Models are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, vol. 33, edited by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, 1877–901.

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. "Named Entity Recognition and Classification In Historical Documents: A Survey." *ACM Computing Surveys* 56, no. 2: 1–47.

Fink, Matthias A., Arved Bischoff, Christoph A. Fink, Martin Moll, Jonas Kroschke, Luca Dulz, Claus Peter Heußel, Hans-Ulrich Kauczor, and Tim F. Weber. 2023. "Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer." *Radiology* 308, no. 3: e231362.

Fullilove, Courtney. 2019. *The Profit of the Earth: The Global Seeds of American Agriculture*. Chicago: University of Chicago Press.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120, no. 30: e2305016120.

Hodge, Walter Henricks, and C. O. Erlanson. 1956. "Federal Plant Introduction—A Review." *Economic Botany* 10, no. 4: 299–334.

Hyland, Howard L. 1977. "History of U. S. Plant Introduction." *Environmental Review* 2, no. 4: 26–32.

Khraisha, Qusai, Sophie Put, Johanna Kappenberg, Azza Warraitch, and Kristin Hadfield. 2024. "Can large Language Models Replace Humans in Systematic Reviews? Evaluating GPT-4's Efficacy in Screening and Extracting Data from Peer-Reviewed and Grey Literature in Multiple Languages." *Research Synthesis Methods* 15, no. 4: 616–26.

Laato, Joonatan, Jenna Kanerva, John Loehr, Virpi Lummaa, and Filip Ginter. 2025. "Extracting Social Connections from Finnish Karelian Refugee Interviews Using LLMs." In Proceedings of the Computational Humanities Research Conference 2024, edited by Wouter Haverals, Marijin Koolen, and Laure Thompson, 662–680. CEUR Workshop Proceedings.

Li, Anny. 2022. "The World was their Garden: Plant Introductions at the USDA, 1898-1984." Master's thesis, Harvard University.

Ma, Zhiwei, Javier E. Santos, Greg Lackey, Hari Viswanathan, and Daniel O'Malley. 2024. "Information Extraction from Historical Well Records Using a Large Language Model." *Scientific Reports* 14, no. 1: 31702.

Moulaei, Khadijeh, Atiye Yadegari, Mahdi Baharestani, Shayan Farzanbakhsh, Babak Sabet, and Mohammad Reza Afrash. 2024. "Generative Artificial Intelligence in Healthcare: A Scoping Review on Benefits, Challenges and Applications." *International Journal of Medical Informatics* 188, 105474.

Polak, Maciej P., and Dane Morgan. 2024. "Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering." *Nature Communications* 15, no. 1: 1569.

Ries, Thorsten, Karina van Dalen-Oskam, and Fabian Offert. 2024. "Reproducibility and Explainability in Digital Humanities." *International Journal of Digital Humanities* 6, no. 1: 1–7.

Schilling-Wilhelmi, Mara, Martiño Rıos-Garcıa, Sherjeel Shabih, Marıa Victoria Gil, Santiago Miret, Christoph T. Koch, José A. Márquez, and Kevin Maik Jablonka. 2025. "From Text to Insight: Large Language Models for Chemical Data Extraction." *Chemical Society Reviews* 54: 1125–1150.

Schroeder, Noah L., Chris Davis Jaldi, and Shan Zhang. 2025. "Large Language Models with Human-in-the-Loop Validation for Systematic Review Data Extraction." Preprint, arXiv:2501.11840.

Stuhler, Oscar, Cat Dang Ton, and Etienne Ollion. 2025. "From Codebooks to Promptbooks: Extracting Information from Text with Generative Large Language Models." *Sociological Methods & Research* 54, no. 3: 794–848.

Törnberg, Petter. 2023. "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning." Preprint, arXiv:2304.06588.

Underwood, Ted. 2025. "The Impact of Language Models on the Humanities and Vice Versa." *Nature Computational Science* 5, no. 9: 695–697.

Underwood, Ted, Laura K. Nelson, and Matthew Wilkens. 2025. "Can Language Models Represent the Past Without Anachronism?" Preprint, arXiv:2505.00030 [cs.CL]. https://arxiv.org/abs/2505.00030.

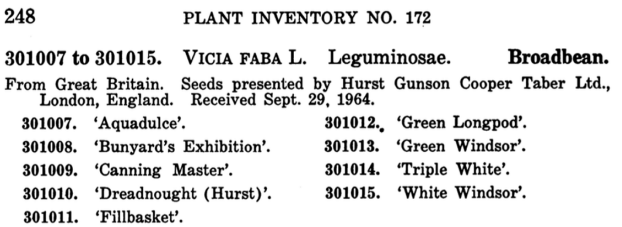
Zhang, Gongbo, Qiao Jin, Yiliang Zhou, Song Wang, Betina Idnay, Yiming Luo, Elizabeth Park, Jordan G. Nestor, Matthew E. Spotnitz, Ali Soroush, Thomas R. Campion Jr., Zhiyong Lu, Chunhua Weng, and Yifan Peng 2024. "Closing the Gap Between Open Source and Commercial Large Language Models for Medical Evidence Summarization." *NPJ Digital Medicine* 7, no. 1: 239.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics* 50, no. 1: 237–91.

Appendix A. Sample input and output from fine-tuned LLM

Example A.1. Example of how the model was able to account for irregular formatting, most notably applying information to a subset of plants.

INPUT:



248 PLANT INVENTORY NO. 172

301007 to 301015. VICIA FABA L. Leguminosae. **Broadbean.**

From Great Britain. Seeds presented by Hurst Gunson Cooper Taber Ltd., London, England. Received Sept. 29, 1964.

301007. 'Aquadulce'.
301008. 'Bunyard's Exhibition'.
301009. 'Canning Master'.
301010. 'Dreadnought (Hurst)'.
301011. 'Fillbasket'.
301012. 'Green Longpod'.
301013. 'Green Windsor'.
301014. 'Triple White'.
301015. 'White Windsor'.

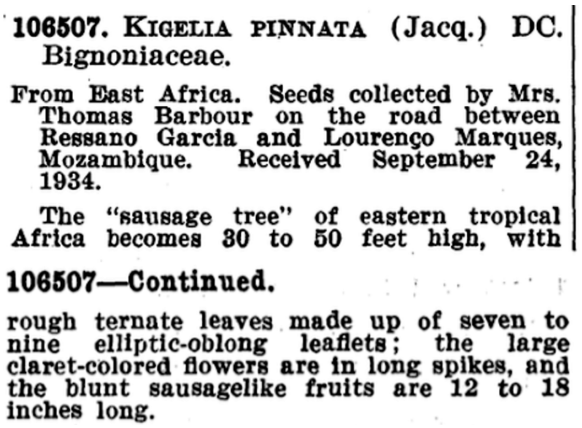
OUTPUT:


```
[
    {
        "pi_number": "301007",
        "name_of_crop": "Vicia faba L. Leguminosae",
        "type": "Broadbean",
        "origin": "Great Britain",
        "received_date": "Sept. 29, 1964",
        "presented_by": "Hurst Gunson Cooper Taber
        ↪  Ltd., London, England",
        "institution_involved": null,
        "experiment_name": null,
        "description": "'Aquadulce'"
    },
    {
        "pi_number": "301008",
        "name_of_crop": "Vicia faba L. Leguminosae",
        "type": "Broadbean",
        "origin": "Great Britain",
        "received_date": "Sept. 29, 1964",
        "presented_by": "Hurst Gunson Cooper Taber
        ↪  Ltd., London, England",
        "institution_involved": null,
        "experiment_name": null,
        "description": "'Bunyard's Exhibition'"
    },

]
```


Example A.2. An example where the model is able to capture seed information that spans multiple pages.

INPUT:



106507. KIGELIA PINNATA (Jacq.) DC. Bignoniaceae.

From East Africa. Seeds collected by Mrs. Thomas Barbour on the road between Ressano Garcia and Lourenço Marques, Mozambique. Received September 24, 1934.

The "sausage tree" of eastern tropical Africa becomes 30 to 50 feet high, with

106507—Continued.

rough ternate leaves made up of seven to nine elliptic-oblong leaflets; the large claret-colored flowers are in long spikes, and the blunt sausagelike fruits are 12 to 18 inches long.

OUTPUT:


```
{
    "pi_number": "106507",
    "name_of_crop": "Kigelia pinnata (Jacq.) DC.
    ↪  Bignoniaceae",
    "type": null,
    "origin": "East Africa",
    "received_date": "September 24, 1934",
    "presented_by": "Mrs. Thomas Barbour",
    "institution_involved": null,
    "experiment_name": null,
    "description": "The \"sausage tree\" of
    ↪  eastern tropical Africa becomes 30 to 50
    ↪  feet high, with rough ternate leaves made
    ↪  up of seven to nine elliptic-oblong
    ↪  leaflets; the large claret-colored
    ↪  flowers are in long spikes, and the blunt
    ↪  sausagelike fruits are 12 to 18 inches
    ↪  long."
}
```


https://doi.org/10.1017/chr.2025.10019 Published online by Cambridge University Press