Unlocking the Past: The Potential of Large Language Models to Revolutionize Transcription of Natural History Collections

Riley Herbst[1,2], Daniel Stille[3], Gilbert F. Gwilliam III[1], Matt von Konrat[1†], Thomas Campbell[3], Wyatt Gaswick[1], Felix Grewe[4], Kimberly Hansen[1], Francisco Iacobelli[5], Karen Jellema[1], Maria Lucia Kawasaki[1], Danny Kreider[1], Richard H. Ree[1], Yarency Rodriguez[1], Alexander Wolpert[2]

[1]Gantz Family Collections Center, Field Museum, 1400 S. DuSable Lake Shore Drive, Chicago, IL 60605, USA

[2]Department of Computer Science, Roosevelt University, 430 S. Michigan Ave, Chicago, IL 60605, USA

[3]Department of Computer Science/Department of Biology, Northeastern Illinois University, 5500 North St. Louis Avenue, Chicago, IL 60625, USA

[4]Grainger Bioinformatics Center, Field Museum, 1400 S. DuSable Lake Shore Drive, Chicago, 60605 IL, USA

[5]Departments of Health Informatics and Data Science and Surgery, Loyola University, 1032 W. Sheridan Rd., Chicago, IL 60660, USA

Keywords: OCR; Digitization; LLM prompt engineering; Biodiversity data; Herbarium vouchers

Citation: Herbst R., Stille D., Gwilliam III G.F., et al.: Unlocking the Past: The Potential of Large Language Models to Revolutionize Transcription of Natural History Collections. Data Intelligence, Vol. 7, Art. No.: 2025r09, pp. 237–264, 2025. DOI: https://doi.org/10.3724/2096-7004.di.2025.0011

ABSTRACT

Globally, natural history collections (NHCs) house nearly three billion scientific collections offering an unparalleled resource that contributes to both science and society. The first two decades of the twenty-first century have seen a rapid rise in the mobilization of digital biodiversity data. Yet, the transcription of important scientific information from NHCs remains a challenge presenting a critical bottleneck in the digitization process. This experimental study investigates the application of Large Language Models (LLMs) in automating the transcription of herbarium specimen labels. Various LLMs were tested, including OpenAI's GPT models and Anthropic's Claude, to evaluate their accuracy and efficiency in transcribing label data. Algorithms have been developed to automate the workflow, leveraging the image processing capabilities of LLMs to bypass traditional Optical Character Recognition (OCR) tools. Prompt engineering strategies were explored to enhance transcription accuracy, focusing on field-specific instructions and formatting. The study demonstrates the potential for LLMs to streamline digitization workflows in natural history collections but underscores several challenges, such as transcription errors, logic errors and inconsistency of results. Promising directions that emerge include improving prompt engineering

† Corresponding author: Matt von Konrat (E-mail: mvonkonrat@fieldmuseum.org; ORCID: 0000-0001-9579-5325).

© 2025 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

techniques, exploring the integration of LLMs into workflows and model selection to achieve higher accuracy and reliability.

1. INTRODUCTION

1.1 Museums and Digitization and Industrial Scale Digitization

Globally, thousands of natural history collections (NHCs) house an estimated three billion scientific collections [1], providing unparalleled resources with broad taxonomic, geographic, and temporal scope that contribute to both science and society, e.g., [2-4]. However, this enormous scientific legacy is largely locked into the typed or handwritten labels mounted with the specimens or stored in associated ledgers and field notebooks [5]. Over the past two decades, there has been a rapid increase in the mobilization of digital biodiversity data [6-7]. For example, the US National Science Foundation (NSF) has provided support for the digitization of approximately 62 million biodiversity specimens from US NHCs [8], enhancing research in global change, biodiversity sciences, and other fields, e.g., [9-11].

1.2 Overcoming Transcription Bottlenecks

The process of digitization encompasses several key tasks [12-13] with one of the most critical being the transcription of specimen label data into electronic formats, often in combination with the capture of a digital image of the specimen [14]. Label data includes vital information on where, when, and by whom a specimen was collected, its taxonomic identification(s), natural history observations of the organism and its habitat and more. Transcribing this information has posed significant bottlenecks [15], generally requiring trained technicians to manually enter data into a dedicated collection management system [5] and recognized as among the slowest steps [16]. Over the past decade, the development of online platforms for crowd-sourced data entry by the general public [17] has helped accelerate transcription. In some cases transcription has also been outsourced to companies that specialize in document transcription [18].

1.2.1 Automated Approaches to Label Transcription

Automation or semi-automation in the transcription pipeline can significantly enhance the speed, accuracy, and cost-effectiveness of data extraction, particularly through technologies like Optical Character Recognition (OCR) and Natural Language Processing (NLP). Automating even the simplest tasks such as sorting labels by language or writing method (typed versus handwritten) stands to improve the overall efficiency of the human-in-the-loop approach [5]. Many platforms and workflows deploying OCR increasing the efficiency of digitization for herbarium specimens have been developed over the years and are in a constant state of advancement, e.g., [1, 19].

An automated system developed by [20] combines OCR with Named Entity Recognition (NER) to enhance the accuracy of label transcription. That system has improved the extraction of structured data from specimen labels, addressing many limitations associated with OCR, such as the misreading of

handwritten or poor-quality prints [21]. Despite these advancements, challenges remain, particularly with complex handwritten labels, necessitating further development of machine learning algorithms to enhance accuracy. In semi-automated workflows, methods like SALIX combine OCR, machine learning, and human oversight to increase the efficiency of label transcription while ensuring quality control [22].

Large Language Models (LLMs), offer new approaches to automating the transcription of herbarium specimen labels and natural history collections, generally. In [23], authors examine the potential of LLMs to enhance herbarium digitization workflows through automated label transcription and developed the platform VoucherVision. In this system, text extracted from herbarium labels via OCR is processed by an LLM, which organizes the data into structured fields and makes it available to human curators for review and validation. However, [23] also highlighted several challenges associated with LLMs at that time, e.g., inconsistent behavior, sometimes generating inaccurate or nonsensical outputs. These "hallucinations" have come to include a vast array of failures of factuality [24]. Another approach to improving the accuracy of automated transcription systems is to combine LLMs with rule-based Natural Language Processing (NLP) techniques. [16] introduced an ensemble approach that combines LLMs with rule-based NLP systems to create a more reliable workflow for herbarium transcription, demonstrating improved accuracy through the integration of multiple machine-learning techniques.

1.3 Goals and Aims

The aim of this study is to assess the accuracy and effectiveness of LLMs in transcribing and managing herbarium voucher data. We design and test methods to measure transcription accuracy, while also evaluating how LLMs can improve transcription efficiency and throughput. We developed accuracy and evaluation tests primarily on typed labels, as handwriting recognition introduces additional variability and challenges that are beyond the scope of this study. In this paper we

1. Develop and evaluate accuracy metrics for LLM-based herbarium label transcription.
2. Create and test automated transcription workflows through:
    - Development of a portable transcription script
    - Direct image-to-text processing (bypassing OCR)
    - Documentation of API costs across LLM services
3. Assess workflow efficiency by:
    - Comparing processing times between automated and manual methods
    - Evaluating prompt engineering strategies
    - Analyzing transcription accuracy through statistical methods

Although our study is restricted to commercially available LLMs, the analysis, pipelines, and prompt engineering strategies we developed are designed to be broadly applicable. Recognizing that accessibility to some commercial LLMs is limited, these methods can be readily adapted for use with non-commercial and open-source LLMs, offering flexibility for individual institutional and digitization needs. Further, this

adaptability ensures that our methods and workflow remain effective in the face of the rapid advancement of LLM technology.

2. METHODS

2.1 Early Pilot Strategies and LLMs

A brief description of early phases that led to the study, as outlined in this paper, is described in Appendix A.

2.2 Software, Tools & Repositories

Appendix B contains a list of software, packages, repositories as well as data aggregators. In addition, access to all Python and R scripts, prompt versions, and ground truth data is available at: https://github.com/rherbst123/FieldMuseumTranscription

2.3 Data Set Description and Definitions of Common Biodiversity Information

Appendix C provides a list of eight data sets utilized for investigating the effectiveness of prompt engineering and LLM accuracy, including a description of the data set (e.g., typed labels, plant group etc.), typical image size and the number of images. The data sets can be accessed at: https://github.com/rherbst123/FieldMuseumTranscription/tree/main/DataAnalysis/DataSets. Multiple plant groups and spore-producing organisms were represented, including flowering plants, ferns, bryophytes, fungi, and lichens. Ground truth data for the images were prepared by collections experts, who transcribed information from the selected images using standard rules. After pilot studies, the initial batch of 300 vouchers was reduced to 100 bryophyte collections. Operationally, plant groups or spore-producing organisms would have different pipelines and would be processed separately from each other. In this study, we chose to focus primarily on bryophyte collections because of the synergy with a large scale digitization effort. Most experiments focused on full-sheet specimen images with typed labels. Some experiments used images cropped to include only the labels.

Early phases of this study experimented with common biodiversity data standards for prompt development. Darwin Core is a community-developed biodiversity data standard that is widely adopted [25]. Darwin Core currently includes over 150 potential fields for biological collections, from which we selected subsets following a hierarchical approach. We initially aimed for broad categories requesting both label information and details that the LLM might infer from the specimen image. This was done through fine tuned prompts (See section 2.5 for details). We further refined our prompts to the LLM, to extract 24 fields, including essential ones like locality that were broken down into 'country', 'state', 'municipality', etc. For accuracy assessment we focused on two evaluation subsets. 1) An almost completely databased label with 17 fields and 2) An abbreviated set of label data composed of five fields, often called skeleton fields representing an essential set of basic data per specimen [26]. This tiered approach

allowed systematic evaluation of LLM performance across different levels of data completeness, from comprehensive documentation to essential skeletal records.

2.4 Pipeline & Portability

The images are downloaded, indexed and paired with the prompt, provided in a text file. Each image is then sent individually to the LLM with the set prompt (Figure 1). The LLM output is converted to a CSV file. That is, we read the formatted output for each image, field by field, and place the information into the appropriate columns. The Python script which handles a run was developed to be portable and to allow for the swapping out of LLMs and adding new LLMs, including non-commercial and open-sourced models. Further, it offers the user the ability to immediately compare transcriptions to ground truth data.

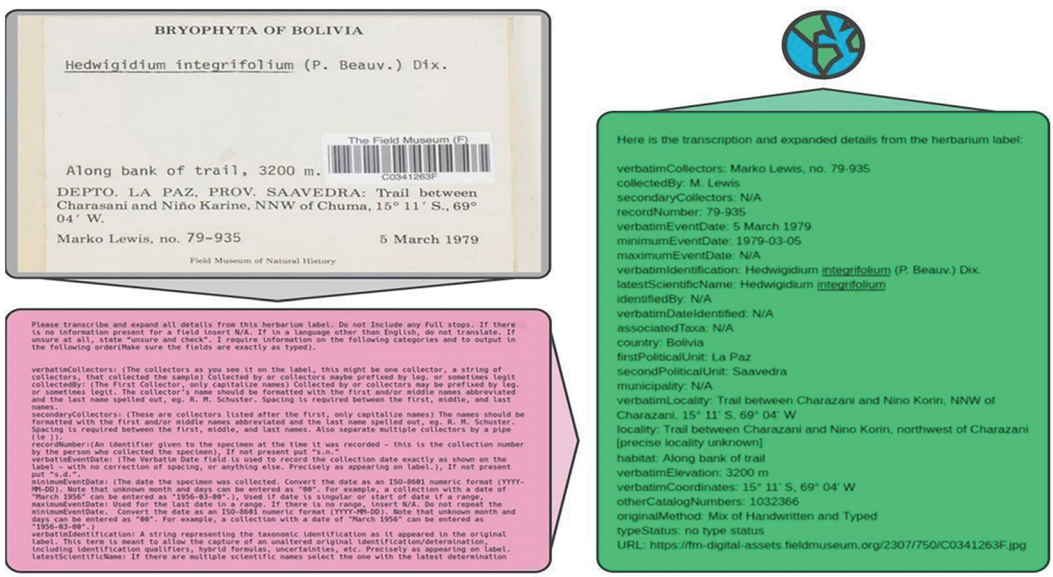


Figure 1. Simplified visual representation of the three core components of a Python script, including an image of a herbarium label, a prompt, and outlet from the LLM into structured text.

2.5 Prompt Engineering Methodology and Validation

We developed and validated our prompt engineering methodology through systematic testing and quantitative evaluation of transcription accuracy. Our approach focused on optimizing three key aspects: field definition clarity, output standardization, and database compatibility. To achieve accurate and consistent transcription of herbarium specimen data, we designed prompts to capture essential fields, validated through iterative testing and alignment checks. Initial prompts focused on categorizing extracted text into key fields (e.g., "Taxon", "Family", "Collector", "Collection Number", "Date", "Locality", "Location", "Habitat"). Early tests revealed that precise, field-specific instructions improved the model's performance in recognizing and organizing label information.

With GPT-4's integrated OCR capability, our workflow evolved to include image-based data extraction, enhancing transcription consistency. We standardized field names using adapted Darwin Core terminology and tailored definitions to improve clarity. For example, renaming 'recordedBy' to 'collectedBy' and including common abbreviations ("leg.," "legit") enhanced the LLM's accuracy in identifying collectors.

2.5.1 Prompt Design Framework

We evaluated prompt effectiveness using three primary metrics: 1) Field-level accuracy, measured by comparison against manually verified transcriptions; 2) Format compliance with aligned LLM outputs to internal database standards, e.g., instructions to expand units (e.g., "m" to "meters"); 3) Consistency of output across multiple runs on the same specimens. Each prompt iteration was tested against a validation set of specimen records, with accuracy measured compared to expert-verified transcriptions.

2.5.2 Prompt Descriptions and Evolution

Early prompts (e.g., Prompt 1.1, Prompt 1.2, Prompt 1.5) were tested interactively with a few voucher images and then modified for later testing with larger batches of images. We include only descriptions of the prompts which were used for large batch tests described in the Results, Section 3.0. Full prompts can be accessed at: https://github.com/rherbst123/FieldMuseumTranscription/tree/main/Prompts

Baseline Performance (Prompt 1.0)

This included a preamble which set out general instructions for the LLM, and definitions of 43 fields, generally following DarwinCore terminology for the definitions. Fields ranged from those associated with higher classification to language to reproductive features.

Optimized Design (Prompt 1.3)

Shortened preamble, and reduced to 25 fields. Some of the field definitions were modified, including more colloquial language rather than the technical language of the DarwinCore definitions, with some adjustments to correct formatting issues.

Format Standardization (Prompt 1.4)

Similar to Prompt 1.3, with additional adjustments to the definitions to exactly match the formatting of some fields (e.g., collectedBy, secondaryCollectors) to the transcription rules used internally.

Final Optimization (Prompt 1.5.2)

Similar to Prompt 1.4, some definitions were adjusted to provide contextual clues for complex fields, e.g., verbatimIdentification and latestScientificName.

2.6 Data Analysis

Our data analysis aimed to assess the accuracy of the LLM transcriptions by comparing them against ground truth data on both small and large scales.

2.6.1 Initial Accuracy Assessment: Manual Comparisons

For early testing with individual images and small batches, accuracy was evaluated manually. We compared the LLM outputs directly with the images on a case-by-case basis to identify any transcription errors and assess prompt effectiveness.

2.6.2 String Matches and Batch Analysis of Transcription Accuracy

For larger data sets, we used the R statistical computing environment to automate comparisons between LLM transcriptions and ground truth targets:

- String Comparison: We used the comparedf function from the arsenal package in R to compare LLM transcriptions of an image with ground truth targets. Other than allowing for case differences and leading and trailing whitespace, matches had to be exact string matches.
- Accuracy Calculation: We calculated the accuracy of an LLM transcript of a dataset as the percentage of LLM transcription values that matched valid ground truth targets (a target that was not "N/A").
- A structured file organized by fields was prepared which listed ground truth data with the LLM output data which did not match exactly. This enabled us to evaluate the types of errors that were occurring in a given field, and provided additional feedback for improving the prompt language.

This initial method offered a rough estimate of transcription accuracy but revealed limitations in precision due to errors where the LLM provided output data other than "N/A" for non-valid targets, leading us to develop an improved approach.

2.6.3 Refined Accuracy Assessment Using Python

To enhance accuracy, we implemented a Python-based analysis pipeline, available at: https://github.com/rherbst123/FieldMuseumTranscription/tree/main/DataAnalysis/AnalysisTools

This refined method categorized ground truth targets as either valid (non-"N/A") or non-valid ("N/A") for each image, enabling more detailed accuracy metrics:

- Match Categories: We compared LLM output values to ground truth target values using a four-way classification:
  - matchValid: Correct matches for valid targets (ground truth is not "N/A").
  - matchNonValid: Correct matches of "N/A" entries.
  - noMatchValid: Valid entries that did not match.
  - noMatchNonValid: Non-valid entries where the LLM output erroneously provided data.

- We focused on matchValid and both types of non-matches (valid and non-valid), as these were most meaningful for assessing transcription accuracy. Non-matches in non-valid entries were especially significant, as they indicated cases where the LLM produced transcription data other than "N/A" when none was expected. Matches in non-valid entries were less important for us as matching "N/A" to "N/A" is a much lower bar than matching string data to string data. Hence we left out matchNonValid counts from our accuracy calculations, i.e., overall accuracy and individual field accuracy where overall accuracy is computed as follows,

$$Overall\ Accuracy = \frac{\#\ matchValid}{total - \#\ matchNonValid}$$

and accuracy for individual fields is determined similarly, using just the counts for that field.

2.7 Glossary of Terms

Appendix D provides a list of key terms used in this paper and their definition, including Ground truth, Levenshtein distance, Field, Value, Target, Transcript, Valid and Overall accuracy.

3. RESULTS

Table 1 summarizes the parameters and results of the series of test runs that were made across LLMs, iterations of prompts, and specimen sets, which are keyed to a reference number (Ref. No.). The results from the various experiments are outlined below. The data for our results is available through this link: https://github.com/rherbst123/FieldMuseumTranscription/tree/main/DataAnalysis

3.1 Prompt Testing and Initial Optimization

3.1.1 Baseline performance (Prompt 1.0; Ref. No. 3.1)

The initial prompt, Prompt 1.0, was designed to extract as much information as possible from each label, including both explicit and inferred data, to evaluate the LLM's full transcription potential. This included a total of 43 fields. In tests with single images and small batches, the LLM showed promising results for simple fields, identifying details like collection dates and taxon names with reasonable accuracy. Additionally, the model demonstrated some inference capabilities by accurately identifying the country when it was not explicit on the label. Similarly, the LLM demonstrated the ability to extract non-label information, e.g., the number of individual specimens on a sheet, and to some degree identifying reproductive features such as flowers on the sheet that were not recorded on the label. However, it often generated hallucinated data, for example, creating species names or localities not mentioned on the label. In one case, it identified the locality as a (real) town in Illinois that was completely unrelated to the actual location of the specimen as described on the label. These hallucinations necessitated scope refinement within the prompt to avoid inaccurate assumptions by the LLM. We added instructions to the prompt

Table 1. A summary providing a reference number (Ref. No.) each referring to a specific testing task, the specific LLM model, the specific prompt version, the number of fields, a description of the specimen set and overall accuracy.

Ref. No.	Task	Model	Prompt version	No. of fields	Specimen set	Overall accuracy
3.1	Full prompt	GPT-4	Prompt 1.0	43	Various	
3.2a	Shortened prompt	GPT-4	Prompt 1.3	25	Data Set 1	43.3%
3.2b	Adjusted prompt: formatting issues	GPT-4	Prompt 1.4	24	Data Set 1	43.8%
3.3	Compare label types: typed	GPT-4	Prompt 1.4	24	Data Set 2	47.2%
3.4a	Replicates, run 1	GPT-4	Prompt 1.5.2	24	Data Set 3	53.3%
3.4b	Replicates, run 2	GPT-4	Prompt 1.5.2	24	Data Set 3	50.5%
3.5a	Original images	GPT-4o	Prompt 1.5.2	24	Data Set 4	57.9%
3.5b	Enhanced contrast	GPT-4o	Prompt 1.5.2	24	Data Set 6	58.4%
3.6a	Full images	GPT-4	Prompt 1.5.2	24	Data Set 5	50.2%
3.6b	Cropped to labels	GPT-4	Prompt 1.5.2	24	Data Set 7	56.7%
3.7a	Replicates with seed and temperature set, run 1	GPT-4o	Prompt 1.5.2	24	Data Set 5	60.9%
3.7b	Replicates with seed and temperature set, run 2	GPT-4o	Prompt 1.5.2	24	Data Set 5	60.5%
3.8a	Test Claude Opus, run 1	Claude 3 Opus	Prompt 1.5.2	24	Data Set 5	40.9%
3.8b	Test Claude Opus, run 2	Claude 3 Opus	Prompt 1.5.2	24	Data Set 5	40.9%
3.9a	Test Google Gemini, run 1	Gemini 1.5 Pro	Prompt 1.5.2	24	Data Set 5	43.9%
3.9b	Test Google Gemini, run 2	Gemini 1.5 Pro	Prompt 1.5.2	24	Data Set 5	44.6%
3.10a	Test Claude Sonnet, run 1	Claude 3.5 Sonnet	Prompt 1.5.2	24	Data Set 5	52.6%
3.10b	Test Claude Sonnet, run 2	Claude 3.5 Sonnet	Prompt 1.5.2	24	Data Set 5	52.4%
3.11a	Non “verbatim” fields	GPT-4o	Prompt 1.5.2	17	Data Set 5	68.4%
3.11b	Skeleton fields only	GPT-4o	Prompt 1.5.2	5	Data Set 5	79.4%
3.12	Cross-Validated values	GPT-4o (Run 1) and Gemini 1.5 Pro (Run 1)	Prompt 1.5.2	24	Data Set 5	84.1%*

Note: Ref. No. 3.12 overall accuracy on a transcript of cross-validated values, 853 valid targets, rather than the 1839 valid targets of a complete transcript.

advising the LLM to limit responses to verifiable information only, which helped reduce-but did not fully eliminate-hallucinations.

3.1.2 Optimized design and abbreviated prompt (Data set 1; Prompt 1.3, 1.4; Ref. No. 3.2 a,b)

Prompt 1.3 was revised to capture a more targeted set of essential fields (25 of 39). Tested on a dataset of 300 images the prompt yielded an accuracy of 43.3%. A review of errors revealed that some stemmed from minor formatting inconsistencies, such as differences in the spacing of initials in collector names (e.g., "R.B. Foster" vs. "R. B. Foster"), as well as transcription errors in longer strings like "verbatimLocality." We modified Prompt 1.4 to include field-specific formatting instructions, raising accuracy of, for example, the collectedBy field from 30.8% to 43.0%.

3.1.3 Abbreviated prompt and reduced data set to typed labels only (Data set 2; Prompt 1.4; Ref. No. 3.3).

For further refinement and to test the impact of label variability, we created a subset of 202 typed-only label images, removing handwritten labels. Applying Prompt 1.4 on this subset improved accuracy to 47.2%. We limited all further accuracy and evaluation tests to typed labels only.

3.2 Image Set Creation, Replication, and Image Processing

3.2.1 Replication of LLM runs (Data set 3; Prompt 1.5.2; Ref. No. 3.4a,b)

We tested for consistency of LLM transcription with the bryophyte data set (Data set 3). Two consecutive runs using the GPT4 model gave somewhat different results. Comparing the two runs with each other, 28.9% of the fields showed a difference. Some of these differences were minor, while others were quite significant. For example, 'verbatimEventDate': "3-11-1981" versus "3-11-1981" (en-dash versus em-dash) and 'verbatimCollector': "Ynes Mexia" versus "Ynes Mexía" (missing accent) are relatively minor: 'verbatimIdentification': "*Fuscoletium dimorphotrichum* (Spruce) Burch" versus "*Vestibulina semibolbosa* (Spruce) Braithw.", or "*Entodon* cf. *nanceclimacium* C.M. (E. e M. Banks 1982.)" versus "*Entodon* cf. *concinnus* (Sull. & Lesq.) Par." are more significant.

3.2.2 Image Manipulation Techniques: Contrast Enhancement and Label Cropping (Data set 4, 6; Prompt 1.5.2; Ref. No. 3.5)

To investigate the impact of image quality, we tested a set of 100 bryophyte voucher images (96 full herbarium sheets, 4 packets) with enhanced contrast. The goal was to improve legibility, particularly for numerals and fine details, which often lead to errors. However, the results showed only a slight improvement in accuracy. A paired t-test of the field-by-field results showed no significant differences indicating that contrast adjustments alone do not significantly affect accuracy for these LLMs. We then tested cropping the image to include only the label area, removing extraneous visual elements from the full herbarium specimen image. This led to a more substantial increase in accuracy on cropped labels, with specific improvements in complex fields like 'locality' and 'verbatimCoordinates'. A paired t-test of the field-by-field accuracies

showed a 15.7% mean improvement in accuracy for the cropped labels (see Table 2), with a p-value of 0.00004. Fields that contain dense text or numerals, benefited most from a focus on text-only data.

Table 2. A comparison of the overall accuracy between a full image and cropped label between ten different fields.

Field name	Overall accuracy: full image	Overall accuracy: cropped label
collectedBy	64.0%	85.0%
recordNumber	53.0%	74.0%
minimumEventDate	69.7%	88.9%
verbatimIdentification	32.0%	46.0%
latestScientificName	50.0%	61.0%
identifiedBy	46.8%	54.6%
firstPoliticalUnit	74.4%	83.3%
secondPoliticalUnit	32.0%	48.0%
locality	8.0%	22.0%
verbatimCoordinates	8.9%	22.9%

Improvement in accuracy was especially notable for locality, which is typically a long field with more opportunities for transcription errors, and 'verbatimCoordinates', a field of mostly numerals and diacritical marks, which are often mistranscribed. The accuracy for many important fields (e.g., 'collectedBy', 'recordNumber', 'minimumEventDate') are all well over 50% exact match for the cropped labels.

3.3 Model Parameter Adjustments and Model Comparisons

3.3.1 Seed and Temperature Effects (Data set 5; Prompt 1.5.2; Ref. No. 3.7)

During this study, GPT-4o added the ability to adjust some of the parameters of the LLM, specifically the "temperature", which affects the "creativity" of the response. Reducing this value is intended to reduce the hallucinations sometimes generated by the LLM. In addition, we could set a "seed" value: keeping it constant from run to run is intended to reduce variations across runs. We tested GPT-4o with the seed parameter set to 42 and the temperature parameter set to minimum to evaluate the accuracy and repeatability compared to GPT-4. Using the same prompt and data set as section 3.4, we obtained higher overall accuracy and better run to run reproducibility. A run-to-run comparison showed 13.4% of the fields with differences, vs. 28.9% of the fields having differences in the section 3.4 runs with GPT-4.

3.3.2 Model Performance across GPT-4o, Claude, and Gemini (Data set 5; Prompt 1.5.2; Ref. No.s 3.8 to 3.10)

Using Prompt 1.5.2, we tested three different models from Anthropic and Google to assess accuracy and consistency across runs. Two consecutive runs with Anthropic's Claude 3 Opus, using a seed parameter of 42 and the minimum temperature, achieved lower overall accuracy than GPT-4o but it produced much lower run-to-run variation. For Google's Gemini 1.5 Pro, two consecutive runs produced higher overall accuracies surpassing Claude 3 Opus. However, run-to-run variation was high at 35.2%. Notably, Gemini 1.5 Pro lacks a user-adjustable seed parameter, and the temperature was set to its minimum. Two consecutive runs with Anthropic's Claude 3.5 Sonnet, again using a seed of 42 and the minimum temperature, yielded higher accuracy with an exceptionally low variation of only 3.9%. This combination of relatively high accuracy and reproducibility suggests that Claude 3.5 Sonnet is well-suited to further prompt refinement for even better performance. Additionally, Claude 3.5 Sonnet is both faster and more economical than Claude 3 Opus, enhancing its appeal for future use.

3.4 Field-Specific Analysis and Cross-Validation Approaches

3.4.1 Exclusion of Verbatim Fields to Improve Consistency (Data set 5; Prompt 1.5.2 ; Ref. No. 3.11a,b)

Verbatim fields were found to contain frequent formatting discrepancies, often affecting transcription accuracy. For example, 'verbatimCoordinates' errors that can be attributed to spacing, missing commas and missing periods account for 50% of mismatches. From a GPT-4o run we analyzed a subset of 17 fields that exhibit lower variability in transcription. This yielded an accuracy of 68.4%. We then tested transcription accuracy on a "skeleton" set of five core fields ('collectedBy', 'collectionNumber', 'minimumEventDate', 'latestScientificName', and 'country'), achieving an accuracy of 79.4%. This subset aligns closely with minimal data requirements for biodiversity information. The field-by-field match is summarized in Table 3.

Table 3. Accuracy for five targeted fields represents a minimum requirement for partial transcription for biodiversity information.

Field name	Percentage accuracy
collectedBy	69.00%
collectionNumber	64.00%
minimumEventDate	80.00%
latestScientificName	77.00%
country	99.00%

3.4.2 Cross-Validation of outputs between Models (Data set 5; Prompt 1.5.2; Ref. No. 3.12)

We assessed the accuracy of field transcriptions that matched across two LLMs. The accuracy of these cross-validated fields between GPT-4o and Gemini 1.5 Pro was 84.1% (Table 1). Accuracy was calculated on 853 shared values out of 1,839 targets in the ground truth dataset. GPT-4o was chosen for its higher overall accuracy, while Gemini 1.5 Pro was selected due to its cost-free access for limited runs.

3.5 Error Classification

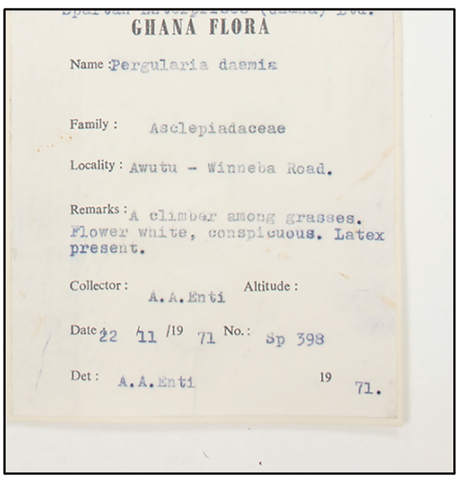
As we evaluated the LLM output from the first few batch tests, it became clear that the errors could be divided into four main categories: transcription errors, logical errors, formatting errors and hallucinations.

Structured files of errors on a field by field basis, including transcription value, ground truth value and a reference (typically a URL) to the voucher image can be accessed at:
https://github.com/rherbst123/FieldMuseumTranscription/tree/main/DataAnalysis/Comparisons/Errors

3.5.1 Transcriptions Errors

Transcription errors involved mistranscribed strings, typically misidentified characters or numerals, such as a "6" for an "8". In two separate runs on GPT-4o with the image on the right, the string "Awutu" was mistranscribed.

Source	locality
gpt4o-2024-06-12-1000	Avutu-Winneba Road
gpt4o-2024-06-12-1500	Avusu Winneba Road
ground truth	Awutu-Winneba Road



3.5.2 Logical Errors

Logical errors are cases when the LLM transcribed the text correctly, but assigned it to the wrong field.

For example, Claude 3 Opus (Ref. 3.8a) assigned the species name "*Macromitrium longicaule* C. Müll." to the field name 'verbatimCollectors', which was present on the label as "A. Touw".

3.5.3 Formatting Errors

As noted in section 3.2 above, many mismatches between the LLM output and the ground truth data were due to formatting which did not match the rules followed by the Field Museum's transcribers, even when the text itself had been accurately transcribed. This type of error is amenable to prompt engineering and is discussed in Section 4.2 below.

3.5.4 Hallucinations

Challenges with hallucinations came in two forms. The first involved false information being transcribed, such as when Claude 3.5 Sonnet (ref. 3.10a) transcribed "E. W. D. Holway" for 'collectedBy', instead of "W. M. A. Brooke". "E. W. D. Holway" appears nowhere on that label, nor does it appear in any ground truth value. The second involved the LLM filling in gaps of information that were not present to begin with. For example in the 'verbatimCoordinates' field, GPT-4o provided minutes in the South coordinates for an image that had minutes only in the East coordinates: transcription "c. 5°41'S 141°05'E" versus ground truth "c. 5° S 141°05'E".

3.6 Accuracy Implications of Prompt Engineering

In early pilot studies we began with a basic prompt aimed at categorizing OCR-extracted text into key fields. We subsequently refined these prompts by adding more precise field definitions. As shown in Table 1 (Ref. Nos. 3.1-3.4), incremental changes improved overall accuracy in GPT-4 runs, although variations across datasets initially obscured these gains. To evaluate accuracy improvements more effectively, we used a consistent dataset of 25 images (dataset 8, a subset of the 100 images in dataset 5, referenced in Ref. Nos. 3.7-3.11) on Claude 3.5 Sonnet. Each prompt iteration is presented in Table 4, showing error reduction and overall accuracy gains as prompts evolved.

Available at:
https://github.com/rherbst123/FieldMuseumTranscription/blob/main/DataAnalysis/Comparisons/BatchComparisons/claude-3.5-sonnet-various_prompts-comparisons.csv

Refinements focused on field-specific formatting and clarity produced the most significant accuracy increase with Prompt 1.5.2, reaching 52.8%. Appendix E outlines specific changes made to each prompt and the corresponding reduction in transcription errors for fields with the highest gains.

Table 4. Comparison of different versions of prompts and overall accuracy.

Prompt	Number errors	Overall accuracy
1.1.1*	286	35.4%
1.2	278	37.1%
1.3	275	37.4%
1.4	297	35.7%
1.5.2	209	52.8%

Note: * Prompt 1.1.1 is our earliest prompt that contains fields consistent with subsequent prompts.

3.6.1 Field-Specific Accuracy Variations Across Models (Data set 5; Prompt 1.5.2; Ref. Nos. 3.7, 3.9, 3.10)

As seen in Table 5, even with the final prompt (1.5.2), field accuracy varied by model. For example, GPT-4o achieved high overall accuracy but did not consistently perform best on every individual field. Such findings underscore opportunities for further refinement and are discussed below.

Table 5. A comparison of field accuracy for 'verbatimCoordinates' and overall accuracy across different LLM models using the same prompt.

Ref. no. (Table 1)	Model	Field accuracy: verbatimCoordinates	Overall accuracy
3.7a	GPT-4o	5.9%	60.9%
3.7b	GPT-4o	2.9%	60.5%
3.9a	Gemini 1.5 Pro	26.5%	43.9%
3.9b	Gemini 1.5 Pro	17.7%	44.7%
3.10a	Claude 3.5 Sonnet	20.6%	52.6%
3.10b	Claude 3.5 Sonnet	17.6%	52.4%

3.7 Cost Performance and Analysis

3.7.1 Time Efficiency: Manual vs. LLM-Assisted Transcription

To assess potential time savings, we compared manual transcription with LLM-assisted workflows on a set of 100 specimens. A collections technician completed transcriptions into a collections management system, EMu, averaging 4.4 minutes per specimen (approximately 13 per hour), while a trained volunteer using Symbiota averaged 2.6 minutes (approximately 23 per hour), benefiting from interface features like drop-down menus. Finally, a simple GUI interface was developed enabling LLM-assisted transcription (using GPT-4o and Prompt 1.5.2). A senior project leader and transcriber was able to review and edit

the LLM-assisted transcriptions at a rate of 1.5 specimen labels per minute (approximately 89 per hour), underscoring the efficiency gains possible with LLM integration, especially for large-scale transcription tasks where minimal post-editing is required.

3.7.2 Cost Implications for Large-Scale LLM-Based Transcription

Using LLMs for large-scale transcription incurs variable costs, depending on the model. Table 6 summarizes costs, processing times, and token usage per specimen across tested LLMs. Of the models tested, GPT-4o offered an optimal balance of accuracy, speed, and cost-efficiency. While Claude models demonstrated high performance, their processing costs were significantly higher. Conversely, Gemini 1.5 Pro is a cost-free option for limited runs but exhibits lower consistency, suggesting it may be suitable for smaller-scale or budget-constrained projects.

Table 6. Performance and cost analysis: LLM cost, time and tokens.

LLM	Unit Cost $	Unit time (sec)	Tokens Used	Notes
GPT-3.5	$0.02	13.3	1505	with Py Tesseract
GPT-4 Turbo	$0.03	11.9	2610	
GPT-4o	$0.02	7.5	3784	
Claude 3 Opus	$0.07	16.6	3208	
Claude 3.5 Sonnet	$0.03	6.9	3208	
Gemini 1.5 Pro	$0.00	6.7	1739	Limited usage

4. DISCUSSION AND FUTURE DIRECTIONS

This study explored the potential of LLMs for automating herbarium specimen label transcription, finding that specific strategies, including prompt engineering, error categorization, image processing, and model selection, substantially influence transcription quality and efficiency. We obtained several key insights into optimizing LLM-based transcription workflows, with implications for both immediate implementation and future development of automated biodiversity data capture systems. Figure 2 outlines a flowchart showing the interconnected stages of LLM-based label transcription.

4.1 Impact of Prompt Engineering and Optimization

Prompt engineering, characterized by three essential components of content knowledge, critical thinking, and iterative design, has rapidly evolved into a crucial field of study, with exponential growth in research publications exploring techniques to optimize model performance (e.g., [27]; [28]; [29]). Recent work by [30] provided a comprehensive review of prompting techniques, emphasizing prompt-based

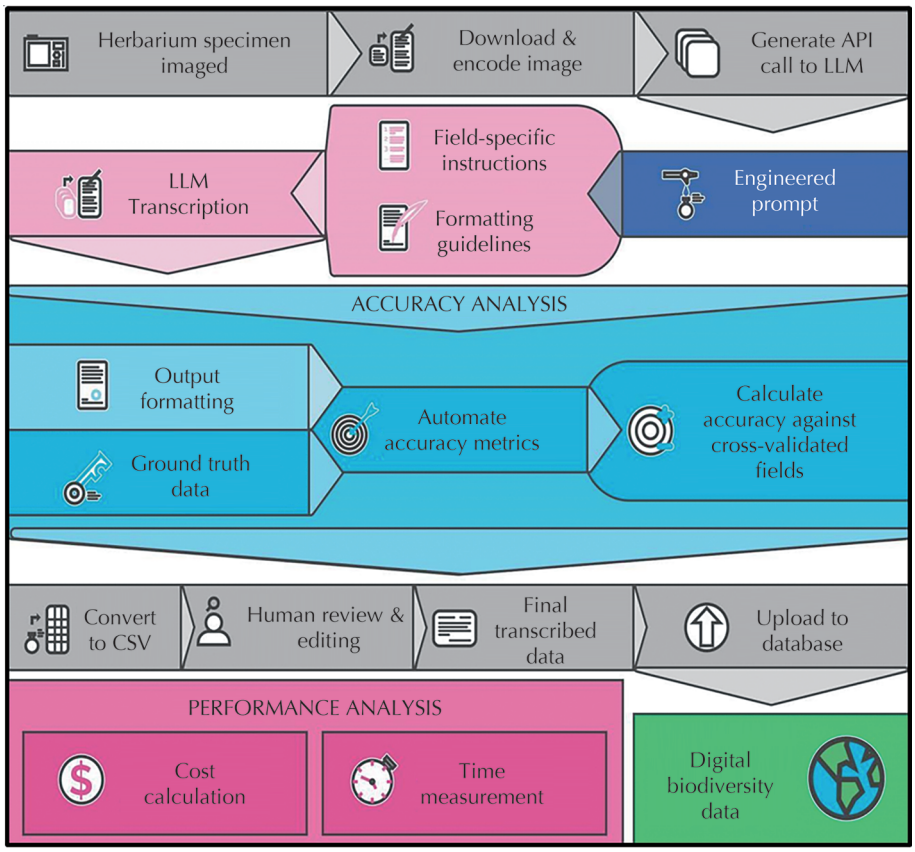


Figure 2. A schema representing a proposed end-to-end workflow described in this paper for automated herbarium specimen label transcription using LLMs, incorporating image processing, prompt engineering, accuracy validation, and performance analysis to transform specimen metadata into standardized digital biodiversity data.

learning as an approach that can significantly enhance output quality. In this study, prompt engineering was critical for improving transcription accuracy across targeted fields (Appendix E). In-Context Learning, where LLMs make predictions based on contexts augmented with a few examples, is sometimes used to guide LLM output without the resource intensive training associated with Machine Learning [23]. Using single-shot learning, where a single example guides the model's output format [31], we achieved significant error reduction in targeted fields. For example, for the fields 'collectedBy' and 'identifiedBy' we reduced our number of errors from 25 to 8 and from 21 to 8 respectively. This approach, applied across GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro, typically achieved 67-80% accuracy for structured fields.

Our results highlight the importance of explicit instruction handling for missing data (e.g., "If no range, insert N/A") which effectively reduced hallucinations, a common challenge in LLMs [24]. Specifying formatting requirements, such as "The collector's name should be formatted with first/middle names

abbreviated and last name spelled out", demonstrated how precise prompt engineering can significantly improve data standardization. These findings contribute to the growing body of work on prompt optimization [30] and suggest that carefully crafted prompts can effectively balance accuracy with resource efficiency in specialized applications like biodiversity data.

Advanced prompting techniques, including bidirectional encoding and Chain-of-Thought (CoT) approaches, also offer promising directions for improvement. This is a common prompting technique [32] where the AI is told to process a prompt in a stepwise manner and illustrate those steps. This could be useful in determining where errors are introduced into a label transcription. For example, Prompt 1.5.2 includes 24 field names and information for how to determine the value for each field. CoT steps could show how accuracy could be improved by changing the field description information in the instruction, or by having the re-reading instruction focus attention on that particular field, perhaps with some additional context. It may be beneficial to elicit Chain-of-Thought reasoning from models that perform relatively accurately in a given field, but less accurately overall. Perhaps CoT reasoning from Gemini 1.5 Pro on how it transcribes the field 'verbatimCoordinates' (see Table 5) could be used to customize a prompt for GPT-4o, which performs more accurately overall.

4.2 Performance Variability

A significant challenge in implementing LLMs is their non-deterministic nature, with run-to-run variability ranging significantly across models. This complicates the process of improving accuracy using prompt engineering, since it is unclear whether changes in accuracy are the result of alterations to the prompt, or unrelated run-to-run variations. This issue is partially mitigated with the introduction of user-tunable parameters that have had a profound role [33] such as "temperature", which allows users to trade off repeatability for "creativity", and "seed", which nominally allows users to set repeatable initial conditions.

4.3 Image Processing and OCR Integration

Our investigation revealed that manual cropping to label-only images improved transcription accuracy by 15.7%, highlighting the importance of focused input data (see Table 2). This suggests that reducing non-relevant visual information allows LLMs to allocate their computational resources more efficiently. Recent advances in vision-language models [34], have introduced new possibilities for OCR-free document understanding, particularly relevant for historical specimen labels. Automated segmentation tools, particularly META's Segment-Anything-Model (SAM) [35], offer promising alternatives for automated label isolation. We found that typed labels consistently yielded higher accuracy than mixed-type labels. Contrast enhancement showed minimal impact on transcription accuracy, suggesting modern LLMs are relatively immune to image quality variations.

4.4 Validation Framework and Standards Integration

Cross-validation between different LLMs emerged as an effective strategy for improving reliability. Given that GPT-4o had the highest accuracy runs and that limited Gemini 1.5 Pro runs incurred no

extra monetary cost, we selected cross-validated data from runs with these two LLMs; accuracy of 84.1% was achieved through cross-validation between GPT-4o and Gemini 1.5 Pro transcripts, though this covered only 46% of all targets. Integration with biodiversity data standards presents opportunities for automated validation against GBIF (Global Biodiversity Information Facility), iDigBio (Integrated Digitized Biocollections) and many other authority sources that are major biodiversity data aggregators that collect, standardize, and make accessible vast amounts of biological specimen data, occurrence records, and taxonomic information from natural history collections and observations worldwide. Our analysis identified limitations in using fixed match thresholds based on Levenshtein distances. For example, LLM transcribed "E. B. Sharp | E. C. Clebsch | K. **I**. **H**ornburgh" compared to ground truth "E. B. Sharp | E. C. Clebsch | K. **R**. **Th**ornburgh", and LLM transcribed "**Ju**lio 3-4, 1977" compared to ground truth "**Jú**lio 3-4, 1977" had the same normalized Levenshtein distances (differences in bold), despite representing significantly different types of transcription variations. This suggests the need for more nuanced, context-aware validation approaches that can distinguish between minor variations and substantial differences in content.

4.5 Practical Implementation and Interface Development

Achieving efficient and cost-effective transcription of label data is a key challenge for digitizing natural history collections [15]. The authors of this paper provided a cost analysis of different approaches to transcribing natural history data, including automated solutions like OCR, manual methods, outsourcing, and crowdsourcing, describing how each has specific benefits and limitations based on institutional resources, goals, and engagement priorities. The application of LLMs introduces a new dimension, balancing cost, accuracy, and speed in real-world settings. Our study has demonstrated that incorporating LLMs into transcription workflows has significantly increased processing efficiency. This efficiency could be further amplified by implementing downstream workflows designed to validate transcription outputs, enabling humans to focus exclusively on more complex or nuanced cases where human oversight is essential for quality control [23]. Such targeted approaches highlight the potential of combining automation with human expertise to optimize large-scale digitization projects. While GPT-4o offers a cost-effective balance of accuracy and speed for large-scale projects, free alternatives like Gemini 1.5 Pro may serve smaller initiatives despite lower accuracy. Future development focuses on creating user-friendly interfaces that eliminate the need for coding expertise, with planned features including simplified API integration and real-time output editing capabilities. The integration of these various approaches ranging from prompt engineering and model selection to validation frameworks and customization strategies, represents a significant step forward in automated herbarium specimen digitization. While challenges remain, particularly in consistency and validation methodology, the rapid advancement of LLM capabilities, combined with increasingly sophisticated implementation strategies, suggests that automated transcription systems will continue improving accuracy and efficiency [23]. As these technologies mature, their integration into institutional workflows could dramatically accelerate the pace of biodiversity data capture, making historical specimen data more accessible to researchers worldwide.

4.6 Model Customization and Fine Tuning

A customized LLM trained specifically to transcribe herbarium specimen labels may produce results that are more aligned with our ground truth targets, but this would require sufficient money, time, and computational resources. [36] discuss the steps that may be taken to customize a locally hosted LLM through fine tuning. The challenge is to imbue the LLMs with domain specific knowledge using optimal allocations of resources. One approach, Parameter-efficient Fine-tuning (PEFT), may be used to reduce training time and memory needs. This efficiency is achieved in part by freezing the first layers of the pre-trained model and retraining the parameters in the last few layers on a dataset specific to a downstream task.

5. CONCLUSION

This study demonstrates that LLMs represent a promising approach for accelerating herbarium specimen label transcription, with highest performance achieved through careful prompt engineering and model selection. Our analysis revealed that GPT-4o delivered the best combination of accuracy (60.9%) and cost-efficiency, while cross-validation between models improved accuracy to 84.1% for matched fields. Several key factors emerged as critical for optimizing transcription: image preprocessing (particularly label cropping, which improved accuracy by 15.7%), model parameter adjustment (especially temperature and seed settings for consistency), and precise prompt engineering (reducing errors by up to 67% in targeted fields). The development of user-friendly interfaces and automated workflows suggests that LLM-based transcription could be readily integrated into existing digitization pipelines. However, challenges remain, particularly in handling handwritten text and ensuring consistent formatting. Future improvements should focus on enhanced validation frameworks, automated image preprocessing, and specialized prompt engineering for different taxonomic groups. While LLMs may still need to achieve the accuracy levels of expert human transcription, they offer a valuable tool for accelerating initial data capture, particularly for typed labels and basic specimen data, potentially transforming the efficiency of large-scale digitization projects in natural history collections. Our study, while focused on commercially available LLMs, presents broadly applicable strategies and workflows that are adaptable to non-commercial and open-source models.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGEMENTS AND FUNDING

The Student Center for Science Engagement at Northeastern Illinois University, the Grainger Bioinformatics Center and the Prince Fellowship (both administered by Field Museum), and the National Science Foundation (Award No. 2001509) provided funding for undergraduate student interns. Financial support was also provided by the National Science Foundation for digitization and curation (Award No.

0949136, 1145898, 1458300, 1541545, 2001509 and 2223875). ChatGPT was also utilized to generate or brainstorm potential coding for Python scripts or potential research directions and testing API payloads. We thank Arianna Salili-James, Ken Norris and Ben Scott of the Natural History Museum, London, for useful discussions regarding mutual interests applying large language models to transcription. We would like to thank the anonymous reviewers and the editorial board of Data Intelligence for the improvement and facilitation of this paper.

AUTHOR CONTRIBUTIONS

Riley Herbst: Conceptualization; methodology; validation; writing review and editing; visualization; software; manuscript preparation and writing original draft.

Daniel Stille: Conceptualization; methodology; validation; writing review and editing; visualization; software; manuscript preparation and writing original draft.

Gilbert F. Gwilliam III: Conceptualization; methodology; validation; writing review and editing; visualization; software; manuscript preparation and writing original draft.

Matt von Konrat: Conceptualization; methodology; data curation; funding acquisition; writing review and editing; writing original draft; project administration; supervision.

Thomas Campbell: Data curation; funding acquisition; writing review and editing; supervision.

Wyatt Gaswick: Conceptualization; data curation; methodology; validation.

Felix Grewe: Conceptualization; methodology; funding acquisition; writing review and editing; project administration.

Kimberly Hansen: Conceptualization; data curation; methodology; validation.

Francisco Iacobelli: Conceptualization; methodology; writing review and editing.

Karen Jellema: Methodology; data curation; writing review and editing.

Maria Lucia Kawasaki: Methodology; data curation; funding acquisition; writing review and editing.

Danny Kreider: Data curation; writing review and editing.

Rick Ree: Conceptualization; methodology; funding acquisition; writing review and editing; project administration.

Yarency Rodriguez: Methodology; data curation; writing review and editing.

Alex Wolpert: Conceptualization; methodology; writing review and editing; manuscript preparation and writing original draft; supervision.

DATA AVAILABILITY STATEMENT

The datasets generated and/or analyzed during this study are available in the Zenodo repository, [![DOI](https://zenodo.org/badge/748320285.svg)] (https://doi.org/10.5281/zenodo.15021660). The data and code can be accessed via https://doi.org/10.5281/zenodo.15021660

SUPPLEMENTARY INFORMATION

APPENDIX A, EARLY PILOT PHASES

Initial tests utilized OpenAI's GPT-3.5 (text-davinci-003) which lacked vision capabilities, necessitating the use of Tesseract OCR for text extraction. OCR outputs were then input into ChatGPT using simple prompts to categorize data into fields such as taxon, collector, and locality. Additionally, various prompt strategies were tested to improve transcription accuracy for individual fields. Evaluation of transcription accuracy by the LLM was limited and ad hoc, but proved viability of the concept. We also considered using Levenshtein distances, a metric used to determine the difference between two strings. However, we decided against using Levenshtein distances in determining accuracy as we saw shortcomings in settling on a threshold value for distances acceptable to a match (see 4.4 Validation Framework and Standards Integration in the Discussion section).

More advanced experiments were performed with OpenAI's GPT-4, which included vision capabilities, so the need for Tesseract OCR was eliminated. In subsequent phases, we also tested other LLMs, such as Anthropic's Claude and models developed by Google, to assess their transcription performance.

APPENDIX B, LIST OF SOFTWARE PACKAGES AND DATA AGGREGATORS

R: 4.4.1; **R Packages: Arsenal**: 3.6.3

Python: 3.11.9; 3.12.5

GIMP 2.10

Field Museum Transcription Python Packages:

Numpy: 1.26.4; **OpenAI**: 1.40.6; **Anthropic**: 0.32.0; **NLTK**: 3.8.1; **Weighted-Levenshtein**: 0.2.2; **Google-GenerativeAI**: 0.7.1; **Requests**: 2.32.3

Graphical User Interfaces (GUI) Python Packages:

Anthropic: 0.32.0; **Numpy**: 1.26.4; **Pandas**: 2.2.2; **OpenCV-Python**: 4.10.0.84; **OpenAI**: 1.40.6; **Tkinter**: Standard Library; **Pillow**: 10.4.0; **Requests**: 2.32.3; **NLTK**: 3.8.1; **TTK Bootstrap**: 1.10.1

Additional Tools & Libraries: Pytesseract: 0.3.13

Biodiversity Data Aggregators:

Images of herbarium voucher specimens were sourced from The Consortium of Bryophyte Herbaria (https://bryophyteportal.org) the Field Museum's online Botanical Collections Database (https://collectionsbotany.fieldmuseum.org)

Github access:

Access to all Python and R scripts, prompt versions, and ground truth data is available at:
https://github.com/rherbst123/FieldMuseumTranscription
An archived and static version can be accessed here:
https://github.com/rherbst123/FieldMuseumTranscription/releases/tag/v1.0

APPENDIX C, OVERVIEW OF EIGHT DATA SETS THAT WERE USED AS PART OF THIS INVESTIGATION

Data Set	Description	Typical Image Size (pixels)	# of Images
1	300 full-sheet and label-only, typed & handwritten labels, multiple plant groups	~1500 x 2250	300
2	Selection of images from Data Set 1 with only typed labels	~1500 x 2250	202
3	Bryophytes from Field Museum collection, only typed labels	~1500 x 2250	78
4	Bryophytes from Field Museum collection, only typed labels	~1500 x 2250	100
5	Bryophytes, only typed labels (different set from Data Set 4)	~1500 x 2250	100
6	Data Set 4 with enhanced contrast	~1500 x 2250	100
7	Data Set 5, cropped to labels only	< ~500 x 500	100
8	Subset of Data Set 5	~1500 x 2250	25

Appendix C provides an overview of eight data sets that were used as part of this investigation. Data set 1 includes a total of 300 herbarium specimen images, of which 280 images were sourced from the Field Museum and 20 images from [37] from several European herbaria. Images included both typed (202) and handwritten or mixed handwritten/typed labels (98) of full standard herbarium sheets and labels only.

APPENDIX D, GLOSSARY OF TERMS

Ground truth: a collection, in spreadsheet form, of data for a set of voucher images. Data for each voucher image has been entered into the spreadsheet by an expert into a set of fields. The data in a ground truth functions as the basis for comparison with LLM output.

Levenshtein distance: a metric for determining the difference between two strings. Levenshtein distance is calculated as the minimum number of edits (insertions, deletions, substitutions) it takes to transform

one string into another. The difference between two strings can be expressed as the raw or normalized Levenshtein distance.

Field: the attribute, or field name, to which data has been applied.

Value: the string data for any given field.

Target: the value for a field in the ground truth for a specific voucher.

Transcript: a full set of transcriptions for a dataset.

Valid: a ground truth target for a field other than "N/A".

Overall accuracy: the number of matches between valid ground truth targets and LLM transcriptions divided by the number of matches between valid ground truth targets and LLM transcriptions plus the number of errors.

APPENDIX E, A SUMMARY OF KEY PROMPT ADJUSTMENTS AND ERROR REDUCTION BY FIELD.

Fieldname	Prompt 1.4 no. errors	Prompt 1.5.2 no. errors	Principal change
collectedBy	25	8	Added more specific formatting information about spacing and initials. Provided example: "eg. R. M. Schuster"
maximumEventDate	20*	0	Inserted: "If there is no range, insert N/A. Do not repeat the minimumEventDate."
verbatimIdentification	19	9	Added "Precisely as appearing on label."
latestScientificName	21	0	Added: "A scientific name is typically a binomial with a genus name and species name." Suggested where to find listed with taxonomic authority or by variety or subspecies
identifiedBy	21	8	Added more specific formatting information about spacing and initials. Provided example: "eg. R. M. Schuster"
firstPoliticalUnit	15	2	Added: "Use the name only and omit the political division tier. This would be immediately after country in terms of hierarchy."
originalMethod	1	5	No change

Note: * for 'maximumEventDate', Prompt 1.4 has the same description as 1.3, yet Prompt 1.3 has 0 errors in this field and there is only one valid target.

REFERENCES

[1] P. W. Sweeney *et al.*, "Large-scale digitization of herbarium specimens: Development and usage of an automated, high-throughput conveyor system," *TAXON*, vol. 67, no. 1, pp. 165–178, 2018, doi: 10.12705/671.10.

[2] C. H. Graham, S. Ferrier, F. Huettman, C. Moritz, and A. T. Peterson, "New developments in museum-based informatics and applications in biodiversity analysis," *Trends Ecol Evol*, vol. 19, no. 9, pp. 497–503, Sep. 2004, doi: 10.1016/j.tree.2004.07.006.

[3] W. Berendsohn, "Using geographical and taxonomic metadata to set priorities in specimen digitization," *Biodiversity Informatics*, vol. 7, pp. 53–62, Oct. 2010, doi: 10.17161/bi.v7i2.3988.

[4] B. Hedrick *et al.*, "Digitization and the future of natural history collections," *BioScience*, vol. 70, no. 3, pp. 243–251, 2020.

[5] D. Owen *et al.*, "Towards a scientific workflow featuring Natural Language Processing for the digitisation of natural history collections," *Research Ideas and Outcomes*, vol. 6, p. e58030, Aug. 2020, doi: 10.3897/rio.6.e58030.

[6] G. Nelson and S. Ellis, "The history and impact of digitization and digital data mobilization on biodiversity research," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 374, no. 1763, p. 20170391, Nov. 2018, doi: 10.1098/rstb.2017.0391.

[7] H. Spiers *et al.*, "Everyone counts? Design considerations in online citizen science," *JCOM*, vol. 18, no. 1, p. A04, Jan. 2019, doi: 10.22323/2.18010204.

[8] J. Lendemer *et al.*, "The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education," *Bioscience*, vol. 70, no. 1, pp. 23–30, Jan. 2020, doi: 10.1093/biosci/biz140.

[9] J. M. Heberling and B. L. Isaac, "Herbarium specimens as exaptations: New uses for old collections," *Am J Bot*, vol. 104, no. 7, pp. 963–965, Jul. 2017, doi: 10.3732/ajb.1700125.

[10] S. A. James *et al.*, "Herbarium data: Global biodiversity and societal botanical needs for novel research," *Appl Plant Sci*, vol. 6, no. 2, p. e1024, Feb. 2018, doi: 10.1002/aps3.1024.

[11] P. S. Soltis, G. Nelson, and S. A. James, "Green digitization: Online botanical collections data answering real-world questions," *Appl Plant Sci*, vol. 6, no. 2, p. e1028, Mar. 2018, doi: 10.1002/aps3.1028.

[12] G. Nelson, D. Paul, G. Riccardi, and A. R. Mast, "Five task clusters that enable efficient and effective digitization of biological collections," *Zookeys*, no. 209, pp. 19–45, 2012, doi: 10.3897/zookeys.209.3135.

[13] G. Nelson *et al.*, "Digitization workflows for flat sheets and packets of plants, algae, and fungi," *Appl Plant Sci*, vol. 3, no. 9, p. apps.1500065, Sep. 2015, doi: 10.3732/apps.1500065.

[14] B. M. Thiers, M. C. Tulig, and K. A. Watson, "Digitization of The New York Botanical Garden Herbarium," *Brittonia*, vol. 68, no. 3, pp. 324–333, 2016.

[15] S. Walton *et al.*, "A cost analysis of transcription systems," *Research Ideas and Outcomes*, vol. 6, 2020, doi: 10.3897/rio.6.e56211.

[16] R. P. Guralnick, R. LaFrance, J. M. Allen, and M. W. Denslow, "Ensemble automated approaches for producing high-quality herbarium digital records," *Appl Plant Sci*, p. e11623, Nov. 2024, doi: 10.1002/aps3.11623.

[17] "Worldwide Engagement for Digitizing Biocollections (WeDigBio): The Biocollections Community's Citizen-Science Space on the Calendar," Laboratory of Molecular Systematics & Evolutionary Genetics. Accessed: Sep. 13, 2024. [Online]. Available: https://www.floridamuseum.ufl.edu/soltis-lab/2018/02/06/worldwide-engagement-for-digitizing-biocollections-wedigbio-the-biocollections-communitys-citizen-science-space-on-the-calendar/

[18] H. Engledow, S. D. Smedt, A. Bogaerts, and Q. Groom, "An Evaluation of In-house versus Out-sourced Data Capture at the Meise Botanic Garden (BR)," *Biodiversity Information Science and Standards*, vol. 2, p. e26514, May 2018, doi: 10.3897/biss.2.26514.

[19] R. Anglin *et al.*, "Improving the Character of Optical Character Recognition (OCR): iDigBio Augmenting OCR Working Group Seeks Collaborators and Strategies to Improve OCR Output and Parsing of OCR Output …," Feb. 2013, doi: 10.9776/13493.

[20] A. Takano, T. C. H. Cole, and H. Konagai, "A novel automated label data extraction and data base generation system from herbarium specimen images using OCR and NER," *Sci Rep*, vol. 14, no. 1, p. 112, Jan. 2024, doi: 10.1038/s41598-023-50179-0.

[21] R. Guralnick *et al.*, "Humans in the loop: Community science and machine learning synergies for overcoming herbarium digitization bottlenecks," *Appl Plant Sci*, p. e11560, Jan. 2024, doi: 10.1002/aps3.11560.

[22] A. Barber, D. Lafferty, and L. R. Landrum, "The SALIX Method: A semi-automated workflow for herbarium specimen digitization," *Taxon*, vol. 62, no. 3, pp. 581–590, 2013.

[23] W. N. Weaver, B. R. Ruhfel, K. J. Lough, and S. A. Smith, "Herbarium specimen label transcription reimagined with large language models: Capabilities, productivity, and risks," *American J of Botany*, vol. 110, no. 12, p. e16256, Dec. 2023, doi: 10.1002/ajb2.16256.

[24] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, Jun. 2024, doi: 10.1038/s41586-024-07421-0.

[25] J. Wieczorek *et al.*, "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard," *PLOS ONE*, vol. 7, no. 1, p. e29715, Jan. 2012, doi: 10.1371/journal.pone.0029715.

[26] M. Tulig, N. Tarnowsky, M. Bevans, A. Kirchgessner, and B. Thiers, "Increasing the efficiency of digitization workflows for herbarium specimens," *ZooKeys*, vol. 209, pp. 103–113, Jul. 2012, doi: 10.3897/zookeys.209.3125.

[27] W. Cain, "Prompting Change: Exploring Prompt Engineering in Large Language Model AI and Its Potential to Transform Education | TechTrends." Accessed: Nov. 03, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s11528-023-00896-0.

[28] Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, Hua Xu, "Improving large language models for clinical named entity recognition via prompt engineering | Journal of the American Medical Informatics Association | Oxford Academic." Accessed: Nov. 03, 2024. [Online]. Available: https://academic.oup.com/jamia/article/31/9/1812/7590607

[29] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications," Feb. 05, 2024, *arXiv*: arXiv:2402.07927. doi: 10.48550/arXiv.2402.07927.

[30] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing | ACM Computing Surveys." Accessed: Nov. 03, 2024. [Online]. Available: https://dl.acm.org/doi/full/10.1145/3560815

[31] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," Jul. 22, 2020, *arXiv*: arXiv:2005.14165. doi: 10.48550/arXiv.2005.14165.

[32] Lance Eliot, "Using the Re-Read Prompting Technique is Doubly Rewarding For Prompt Engineering," *Forbes*, Jul. 06, 2024. [Online]. Available: https://www.forbes.com/sites/lanceeliot/2024/07/06/using-the-re-read-prompting-technique-is-doubly-rewarding-for-prompt-engineering/#open-web-0.

[33] Z.-H. Qiu, S. Guo, M. Xu, T. Zhao, L. Zhang, and T. Yang, "To Cool or not to Cool? Temperature Network Meets Large Foundation Models via DRO," Jun. 16, 2024, *arXiv*: arXiv:2404.04575. doi: 10.48550/arXiv.2404.04575.

[34] G. Kim, S. Yokoo, S. Seo, A. Osanai, Y. Okamoto, and Y. Baek, "On Text Localization in End-to-End OCR-Free Document Understanding Transformer Without Text Localization Supervision," in *Document Analysis and Recognition-ICDAR 2023 Workshops*, vol. 14193, M. Coustaty and A. Fornés, Eds., in Lecture Notes in Computer Science, vol. 14193. , Cham: Springer Nature Switzerland, 2023, pp. 215–232. doi: 10.1007/978-3-031-41498-5_16.

[35] A. Kirillov *et al.*, "Segment Anything," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026. Accessed: Nov. 03, 2024. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html

[36] V. Kushala, W. Harikrishna, and G. Yogesh, "Fine Tuning LLM for Enterprise: Practical Guidelines and Recommendations," Mar. 23, 2024, *arXiv*: arXiv:2404.10779. Accessed: Sep. 09, 2024. [Online]. Available: http://arxiv.org/abs/2404.10779

[37] M. Dillen *et al.*, "A benchmark dataset of herbarium specimen images with label data," *Biodiversity Data Journal*, vol. 7, 2019, doi: 10.3897/BDJ.7.e31817.

AUTHOR BIOGRAPHY

Riley Herbst: Riley Herbst is a volunteer in the Botanical Collections at the Field Museum in Chicago, Illinois. He has a Bachelor's in Computer Science and Mathematics, and a Master's in Computer Science.

Daniel Stille: Daniel Stille is a graduate of Northeastern Illinois University, where he studied Computer Science and Linguistics. His internship and volunteer work has focussed on evaluating LLM performance in both testing phases and within workflows.

Gilbert F. Gwilliam III: G.F. (Jeff) Gwilliam is a volunteer in the Botanical Collections at the Field Museum of Natural History in Chicago. He has an undergraduate degree in physics, and a master's degree in Materials Science and Engineering. His technical career was primarily in semiconductor materials and devices.

Matt von Konrat: Matt von Konrat, Head of Botanical Collections at Field Museum, Chicago specializes in the systematics of bryophytes, particularly liverworts. He is actively engaged in community science initiatives connecting the general public to natural history collections. Recently, von Konrat has developed an interest in applying artificial intelligence to help accelerate biodiversity discovery and documentation.
ORCID: 0000-0001-9579-5325

Thomas Campbell: Thomas Campbell is a senior instructor of biology at Northeastern Illinois University in Chicago, IL, USA. His research interests are in plant biology, genetics, and citizen science projects (people powered science). He is a recipient of the NEIU Instructor Excellence award.
ORCID: 0000-0003-4228-9685

Kimberly Hansen: Kimberly Hansen works for the Field Museum of Natural History as the Collections Manager for Seed Plants and Economic Botany in the Searle Herbarium. Her research interests include floristics and plant systematics.

Francisco Iacobelli: Dr. Francisco Iacobelli's research interests are at the intersection of human-computer interaction and artificial intelligence. This involves structuring unstructured information with and without the use of large language models, and it also involves purposefully designing models of language and behavior for marginalized populations. He is an associate professor with a dual appointment to the departments of Health Informatics and Data Science, and the department of Surgery at Loyola University.

Alex Wolpert: Alexander Wolpert is a full professor of AI and Algorithms at the Department of Computer Science at Roosevelt University. His research interests are in the areas of data representation, interpretable AI, and algorithms for combinatorial problems.