FEDERAL RESERVE BANK PHILADELPHIA
Consumer Finance Institute

Discussion Papers

Combining AI and Established Methods for Historical Document Analysis

Dan Moulton
Federal Reserve Bank of Philadelphia
Consumer Finance Institute

Larry Santucci
Federal Reserve Bank of Philadelphia
Consumer Finance Institute

Robyn Smith
Federal Reserve Bank of Philadelphia
Consumer Finance Institute

DP 25-02

PUBLISHED
November 2025

FRB CFI PHILADELPHIA

DOI: https://doi.org/10.21799/frbp.dp.2025.02

Combining AI and Established Methods for Historical Document Analysis

Dan Moulton†
Daniel.Moulton@phil.frb.org

Larry Santucci†
Larry.Santucci@phil.frb.org

Robyn Smith†
Robyn.Smith@phil.frb.org

October 20, 2025

Abstract

This paper examines methodological approaches for extracting structured data from large-scale historical document archives, comparing "hyperspecialized" versus "adaptive modular" strategies. Using 56 years of Philadelphia property deeds as a case study, we show the benefits of the adaptive modular approach leveraging optical character recognition (OCR), full-text search, and frontier large language models (LLMs) to identify deeds containing specific restrictive use language—achieving 98% precision and 90–98% recall. Our adaptive modular methodology enables analysis of historically important economic phenomena including restrictive property covenants, their precise geographic locations, and the localized neighborhood effects of these restrictions. This approach should be easily adaptable to other research involving deeds and similar documents.

Keywords: large language models (LLMs), artificial intelligence (AI), machine learning (ML), restrictive covenants, deeds, property, real estate, housing, John Coltrane, digitization

JEL Classification: C81, N32, R31, R38

† Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106-1574.

Executive Summary

This paper examines methodological approaches for extracting structured data from large-scale historical document archives, comparing "hyperspecialized" versus "adaptive modular" strategies using 56 years of Philadelphia property deeds as a case study. Hyperspecialized approaches rely on custom models optimized for specific tasks, potentially achieving superior accuracy but requiring substantial investment while offering limited flexibility. Adaptive modular approaches combine existing off-the-shelf or lightly customized tools in ensemble to balance accuracy, cost-effectiveness, and flexibility.

Our case study using Philadelphia deeds shows the benefits of the adaptive modular approach in practice. We use optical character recognition (OCR), full-text search, and frontier large language models (LLMs) to identify deeds containing specific restrictive use language—achieving 98% precision and 90–98% recall. We also demonstrate how our approach proves particularly valuable when research questions evolve organically, as modular components can be repurposed across multiple projects without (re)building specialized models.

We find that LLMs can be particularly valuable in the context of a modular approach. While LLMs can be difficult to control and very costly to run on large amounts of text, they work well in conjunction with other tools. Cost issues can be avoided by focusing LLM tasks on a small subset of selected text. Limiting the scope of text evaluated by LLMs improves their performance and therefore requires less time for human review. We also find that frontier LLMs are very good at overcoming even severe OCR issues—much as human readers can.

Our adaptive modular methodology enables analysis of historically important economic phenomena including restrictive property covenants, their precise geographic locations, and the localized neighborhood effects of these restrictions. This approach should be easily adaptable to other research involving deeds and similar documents.

1 Introduction

For thousands of years, humans have produced vast archives of unstructured textual data that remain largely inaccessible to most contemporary researchers. This unstructured textual data include, but are certainly not limited to diary and journal entries; meeting minutes; organization bylaws; individual and business ledger books; property transfers; birth, death, and marriage certificates; records of immigration and naturalization; shipping and business inventories.

Many administrative documents, including "public records"—such as property deeds or mortgage documentation—have been "digitized" in the sense of scanning or digital imaging. However, this format is unsuitable for large-scale analysis. Gathering, yet alone analyzing, data from millions of pages of text isn't feasible as a manual task and must be tackled technologically.

Advances in optical character recognition (OCR), full-text search, and large language models (LLMs) offer unprecedented opportunities to unlock insights from the historical source material by making it economically and technically feasible to extract, process, and analyze these data. The availability of this new technology and the relative costs and benefits associated with each raises questions about usage and applicability that we tackle by considering the trade-off between deep specialization for a specific use case ("hyperspecialization") versus an approach based on an ensemble of flexible tools that can scale across multiple research questions ("adaptive modularity").

Founded in 2023, the Philadelphia Federal Reserve's Center for the REstoration of Economic Data (CREED[1]) works to harness, refine, and customize these advancements to produce high-quality, publicly available data, code, and economic research. In this paper, we discuss CREED's approach to data extraction and many of the thorny issues that arise. We compare the advantages and disadvantages of two approaches to converting historical source data into a research dataset—hyperspecialization and adaptive modularity. Finally, we provide a case study of our work with 56 years of Philadelphia property deeds (1910–1965) demonstrating how the flexibility of our adaptive modular approach harnesses the most appropriate and cost-effective tools for the job.

2 Hyperspecialization or Adaptive Modularity?

Long histories of administrative data are often available only in physical form—typically paper or microfilm. One approach to extracting data from these media is to go down the path of *hyperspecialization*—building custom, highly tuned machine learning (ML) models that can achieve results that meet or exceed those produced

[1]An homage to Apollo Creed and Rocky Balboa, two of Philadelphia's cinematic heroes.

when humans manually search the documents and transcribe and format the relevant information. Hyperspecialization can be the best path—if the research question is very well defined[2] or the underlying data are highly uniform in terms of font, handwriting, page layout, and image quality.

That said, hyperspecialization is not without its disadvantages. Building custom models incurs costs in multiple ways:

1. Development time: This encompasses the process of writing the relevant code and researching and testing the best models for the task.

2. Training time: The time it takes for a model to be "trained" (i.e., for the computer to do the computations necessary to build and optimize the model). This can be substantial depending on the complexity of the model and/or the size of the dataset.

3. Acquiring training data: Training data are a selected subset of the real source material where the data extraction process is completed manually by humans.[3] The labels[4] applied by humans serve as an answer key to "teach" a model what the data look like, what information to extract, and how the information should be formatted for output after extraction. The performance of a model is highly dependent on the quality of training data. Biases in the training data can propagate into model outputs. Heterogeneities in the data also need to be sufficiently represented[5] in the training data for the model to work well across the entire dataset. Creating high-quality labeled training data requires substantial effort, particularly for historical sources where overall trends and heterogeneities in the full sample are likely not yet well understood at the beginning of a project since they were previously in inaccessible formats.

4. Computing costs: Training a custom model can require more computational power (i.e., additional cores or GPUs) than available via a typical desktop computer. Power can be gained via in-house high-performance computing environments or by renting compute power via cloud services.

[2]Think, for example, of a credit scoring model to predict default risk or a model used to predict customer response to a marketing campaign.

[3]Synthetic data are a promising solution to address bias, cost, and privacy concerns that arise when constructing real-world training data. Most research and success in this area are focused on synthetic image generation for computer vision models. Synthetic text training data presents additional challenges and is, so far, less fruitful. See, e.g., Aranjuelo, García, Loyo, Unzueta, and Otaegui (2021); Bauer et al. (2024); Li, Zhu, Lu, and Yin (2023); Liu et al. (2024).

[4]A "label" is qualitative or quantitative information assigned to part or whole of an image or piece of text. A label could be "cat" for an image of a cat or "3" if it's an image of 3 cats. A label for a paragraph could be "happy" for the tone, or if it's dialogue, the label could be the name of the character speaking. It can also be exactly transcribed data for the purpose of training OCR models.

[5]"Sufficiently represented" is an active area of ML research and is not well defined or understood.

5. Service costs: Frontier[6] LLM models typically charge per input *and* output token.[7] These costs can accumulate quickly when working with large amounts of text. Output token usage is not easy to predict—generally there is not much to constrain the length of an LLM's response[8]—so there is some chance of cost overruns when scaling up small tests to apply LLMs to a full dataset.

These costs tend to increase if the input data are not highly uniform—heterogeneous inputs require more training data[9] and more complicated models, and may increase the risk of overfitting.[10]

Hyperspecialization is, by its very nature, inflexible, which often conflicts with the practical demands of research. Projects can evolve quickly as more is learned about the data, particularly with large new datasets where the contents have never been studied or accessed before. Refining subsequent iterations in training data can lead to the discovery of something that wasn't known to exist or—worse—that invalidates the initial research design. Moreover, even when projects evolve according to plan, researchers often identify additional research questions that may inform future research. If the new research question is sufficiently different from the initial question that the model was trained on, the highly tuned machine must then be rebuilt from scratch—at additional cost.[11]

Alternatively, an adaptive modular approach employs flexible, adaptive tools that achieve high—but not the highest possible—accuracy while allowing a project to evolve organically. This approach utilizes off-the-shelf technology to answer initial research questions or develops new but generalized tools to facilitate a more thorough understanding of the data. Custom models can also be incorporated—typically those that are less sensitive to the size and construction of training data. The underlying components of this modular approach can then be applied to many other research projects.

We also note that these approaches are not mutually exclusive and may indeed

[6] "Frontier" is a common moniker for "cutting-edge" or top-tier LLM models—excluding things like older open source or older closed source models.

[7] A token is similar to, but not exactly the same as, a word. Very roughly, a token averages four characters including spaces.

[8] Often a hard cap on output tokens per query can be set but negatively impacts LLM performance—it's easy to chop off information you need from an LLM response.

[9] Generally speaking, each heterogeneity needs to be appropriately represented in the training data.

[10] Overfitting occurs when a model performs excellently on the training data (which the model has seen as part of training or even evaluation) but poorly on data it has never seen before. Multiple factors can contribute to overfitting, including model complexity and poor construction of the training data, in one or more ways.

[11] This incremental cost can be nearly equal to the initial investment if the research question or machine learning task is sufficiently different than originally conceived. For example, imagine that your research question involves finding all the proper names in a collection of documents and later you decide you also need all addresses. This now requires a new set of training data focused on addresses—text and where that text is on the page.

be complementary. An adaptive modular process can—and often does—help identify where and how the investment in a more sophisticated or costly, targeted, and customized approach is likely to yield a worthwhile return. This iterative approach can minimize the risk of overengineering at too early a stage of data exploration while still allowing for investment in specialized tools at points when the return is greatest.

3 Our Adaptive Modular Approach

CREED's interest in historical data is varied and far-reaching. Our broad research scope favors adaptive and modular approaches, complemented by strategic deployment of hyperspecialization. This approach best supports CREED's mission to facilitate the development of tools that can be used by the broader research community in addition to compiling high-quality, ready-to-use datasets.

Our adaptive modular approach to extracting relevant, structured data from images of textual data starts with scanned images of microfilm or paper records ("input data"). Working with these images can be challenging. Paper records may be torn, faded, creased, or otherwise damaged. The quality of the microfilm image can be dependent on the technician operating the machine—images may be crooked, cut off, blurry, under- or over-exposed, pages may be skipped in multi-page documents, etc. Documents may have been handwritten by one person or by many people, each with their own unique handwriting. Documents may have been typewritten by one person using the same typewriter or multiple people using different typographical conventions with multiple different typewriters, each of which can have unique keyboard structures or other calibration quirks. Finally, these historical documents contain spelling, grammatical, and other types of errors innately expected of humans.

We accept that off-the-shelf OCR quality will be highly variable and often poor given the above described characteristics of the input data. To partially offset these quality issues, we implement flexible, fast fuzzy search[12] to roughly extract data of interest. Once the corpus has been winnowed down to a smaller size, we can efficiently pursue a more hyperspecialized approach. We can then apply custom ML or frontier LLM models to further clean the data. We find that frontier LLMs can correctly comprehend even very garbled OCR output, and that other older ML methods struggle without accurate OCR. Even new but small language models or older versions of frontier LLMs do not handle garbled OCR well. Throughout this process, we ensure that there is always a

[12]A fuzzy search can accommodate spelling differences, i.e., searching for "Jon Smit" will return results with "John Smith."

"human in the loop"[13] (Meng, 2023) by building tools to make it easy to check interim steps[14] and tweak the pipeline as needed. We always validate a large enough sample of results by hand to quantify the accuracy of the broader sample.

We illustrate our adaptive modular approach by applying it to a challenging real-world dataset: 56 years of property deeds from the City of Philadelphia. We begin by discussing the broader value of property deed records before diving into our specific application.

4 Case Study: Philadelphia Property Deeds

4.1 Background on Deeds

The legal process of recording property transfers generates a particularly rich administrative dataset. Deeds have been used to study a wide range of economic, legal, and historical subjects. Key examples include the long run effects of tangled titles[15] (Hincken, 2021), property transfer under intestacy[16] (Stein & Carpenter, 2022), building development and use history, genealogy, environmental history, and the concentration of and pattern of land ownership (Storey, 2018). These topics provide key insight into the functioning of the real estate market—a large driver of the US economy—and therefore into inflation, employment, and financial stability for the economy as a whole.

An additional important topic of ongoing research is the impact of historical zoning practices on housing access, affordability, and quality (Asquith, Mast, & Reed, 2023; Bartik, Gupta, & Milo, 2025). Before the legal ascendancy of zoning, properties were subject to legal covenants that performed a similar, privately initiated, function to zoning (Korngold, 2001; Weiss, 1987). Covenants in property deeds may restrict use, ownership (Santucci, 2020; Surani et al., 2025), occupancy, and a variety of building characteristics, such as height, setback, and exterior finish. To measure the long-term effects of these types of restrictions, researchers must identify and extract the relevant restrictive language and property characteristics and geographically locate each property—all from information contained within the deeds themselves.

Modern versions of property transfer data—what might be found on a municipal

[13] Meng defines human-in-the-loop as the integration of human interaction, intervention, and judgment to control or alter automated processes, a practice increasingly emphasized in ML and AI systems.

[14] We follow the maxim of "ABC—Always Be Checking." We build tools to make quick visual and mathematical checks of our results. This allows us to tweak our processes and to build a feedback loop to optimize tool and model performance.

[15] Tangled titles, also referred to as "heir's properties," describe properties where the name of the owner on the deed is not the same as the apparent owner of the property. Tangled titles are a major issue impeding households from benefiting from property ownership.

[16] Intestacy occurs when an owner of a property dies without having written a will. Property transfer under intestacy is a major cause of tangled titles.

open data portal—contain some of the information necessary to facilitate the research mentioned above. Such databases typically include only the last 15–25 years of records and only select information, not the records in their entirety. For example, in Philadelphia, property transfer data from 1999 onward are publicly available as tabular data but track only price, address, and buyer and seller names. In contrast, existent property records in Philadelphia extend back to the city's establishment in 1682 (Hughes, 1958), nearly all of which are already available on microfilm.

4.2 Adaptive Modular Approach Applied to Philadelphia Deeds

Our exploration and analyses of Philadelphia property deeds require us to examine a wide array of text and transaction types in our corpus of deeds. For example, when identifying cases in which some language of interest exists (positive cases), we may also need to capture a set of deeds that do not contain the item of interest (negative cases). Negative cases can be used as control variables or as counterfactuals for causal analysis. In addition, we want to extract tabular data from the deeds—for example, seller, buyer, mortgage terms, covenant terms, or encumbrances. Finally, we want to pinpoint the geographic location of the deeds of interest, ideally at the parcel level.

Through an ongoing partnership with the Philadelphia Department of Records (DOR), CREED obtained digital images of property deeds from 1910–1965. This dataset contains 4.7 million images representing at least 7 million pages and 2.4 million deeds.[17] Philadelphia's deeds were recorded in three primary formats over time: 1) transcribed by hand, 2) transcribed by typewriter, and 3) photostats[18] made of the original deed documents. We show examples of these formats in Figures 1, 2, and 3 below. The quality of the image captured in the microfilm varies substantially, ranging from exceptionally clear to completely illegible. Variable document format and poor-quality images are likely to be typical in other cities as well, particularly for anything recorded prior to the mid-20th century. Our approach to these problems should be adaptable to many other researcher's use cases.

[17]The microfilm itself is old, produced by an initiative launched in 1949 (Hughes, 1958). The images capture the deeds in a continuous stream, varying between one and two pages per image, with no digital demarcation between documents. For this reason, we report image counts, not pages, and an approximate count of deeds. Discussion of methods for dividing images into individual deeds is beyond the scope of this article.

[18]The photostat machine was an early photocopier that used camera film instead of toner, producing white-on-black negative copies.

Figure 1: Example of a handwritten deed transcription found in Philadelphia prior to approximately 1913. Note the similarity in structure to the typewritten deed transcription in Fig. 2.

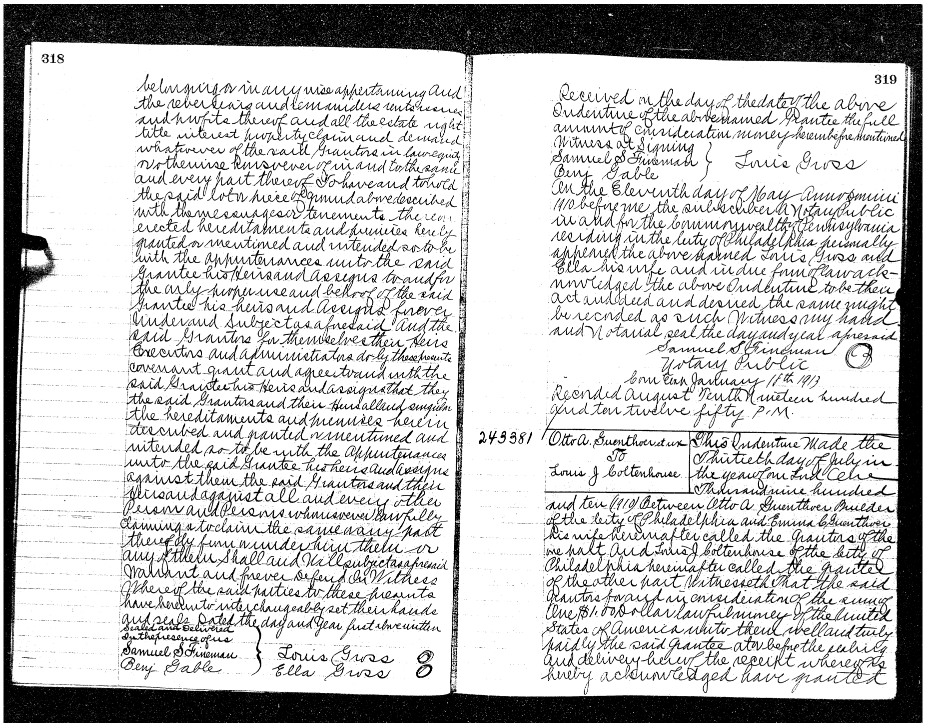


Figure 2: Example of a typewritten deed transcription found in Philadelphia from approximately 1913–1935. Note the similarity in structure to the handwritten deed transcription in Fig. 1.

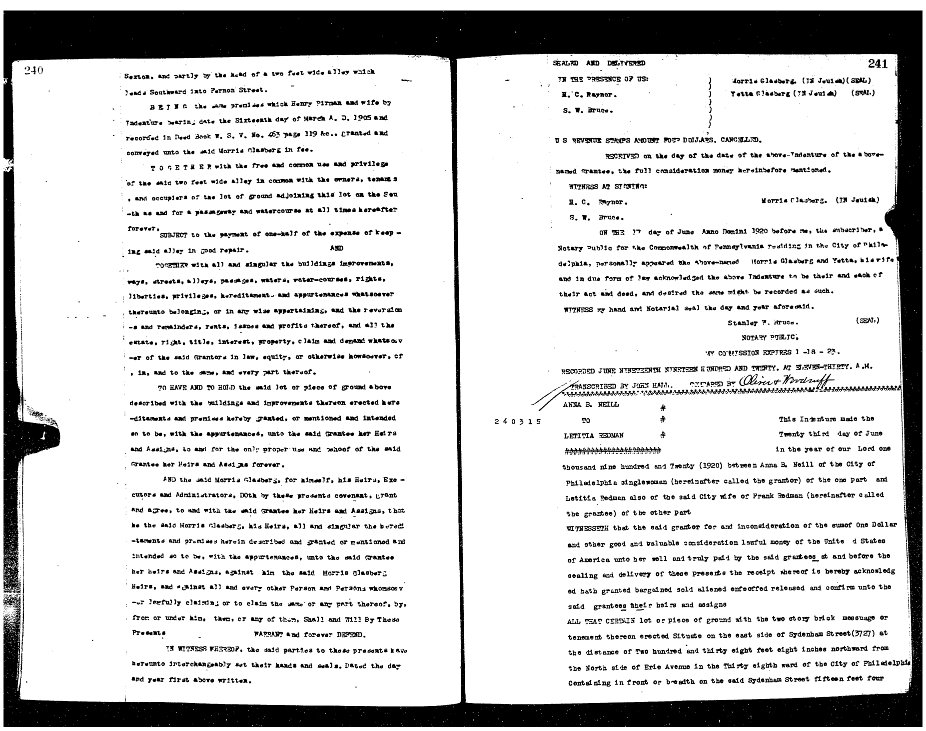
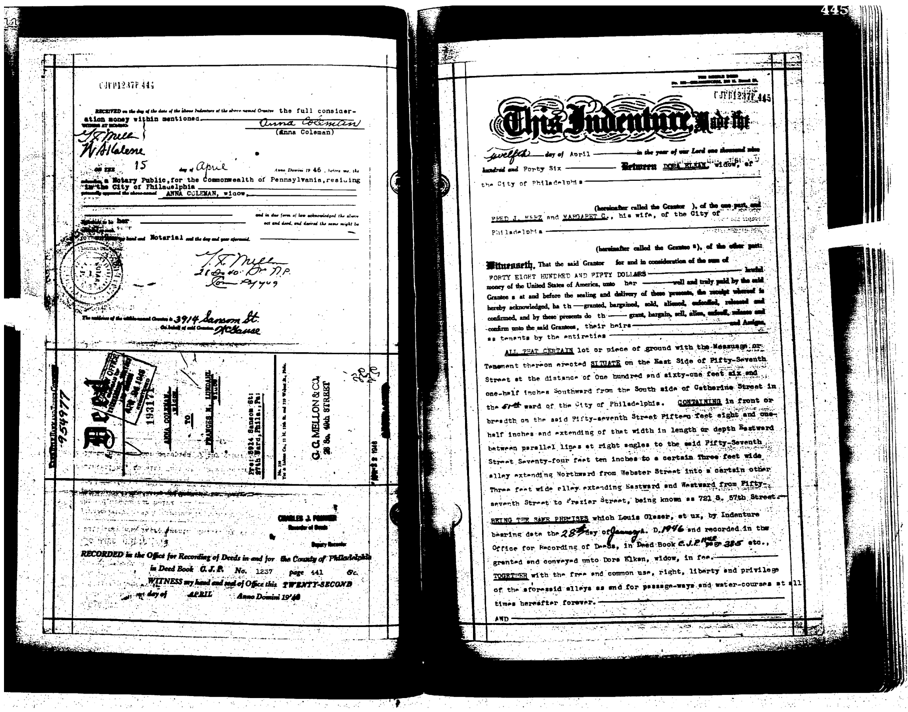


Figure 3: Example of a deed photostat found in Philadelphia after approximately 1935. Note the variety in fonts (both typeset and handwritten), font sizes, and the section of text on the left page that is rotated 90 degrees.



We begin the process of analyzing the property deeds by converting the image files we received from the DOR into textual data using OCR.[19] A custom, fine-tuned OCR model would be particularly challenging, given a wide range of fonts, handwriting, document formats, and image quality. Instead, we leverage off-the-shelf software, in this case Amazon Textract, which offers a balance of scalability, ease of use, cost, and flexibility (i.e., performance across a wide range of inputs).[20]

Unfortunately, this approach produces messy results. Even text clearly legible to humans can be mangled by the OCR process. If this process did read the text perfectly, we could do a simple search for exact terms of interest. For example, for intestate property transfer, we might search for "intestate" or "without will" and be confident that the search results are true positives. However, both false positives and false negatives are introduced by the OCR process. The word "intestate" could be read as "inte3tate"—a false negative with a simple keyword search, or "interstate" could be read by OCR as "intestate"—a false positive. We therefore need to use technology that will allow us to cast a broad, fuzzy net. For example, we may want to search for the phrase "shall not be sold," allowing for other words to be inserted between the search terms (slop)

[19]This process is akin to converting an unsearchable pdf file into a searchable one.

[20]This is not an endorsement of Textract in particular. There are many other viable choices for off-the-shelf OCR, including but not limited to Google Cloud Vision or Azure Computer Vision or free-to-use options such as docTR OCR, EasyOCR, or Tesseract

or allowing variable Damerau–Levenshtein (DL) edit distances.[21]

The particular search method used should be both flexible and fast, allowing for fuzzy search, phrase reordering, and other natural language processing techniques to facilitate balancing recall and precision.[22] To this end, we employ OpenSearch,[23] a full-text[24] search engine that can be found "under the hood" powering a wide variety of consumer products and research programs. NASA uses OpenSearch to provide public search access to Earth Observing System Data.[25] E-commerce websites use it as a search engine for customers to find products. A wide range of companies use it to search log files or as an analytic backend, which can be used to identify and troubleshoot website or software issues, study consumer use pattern and purchase habits, and power business analytics dashboards, amongst many things. In our research setting, OpenSearch allows us to overcome some of the issues with OCR quality while avoiding the process of constructing an expensive, custom OCR solution. OpenSearch offers additional advantages, for example, it allows easy access to the underlying data in a less technical, more intuitive way. Just as anyone can search an e-commerce website, a researcher studying Philadelphia deeds can search our corpus of deeds for a variety of research targets, including language banning the selling of "spirituous liquors," finding foreclosed properties, or finding the deeds associated with jazz great and Philadelphia resident John Coltrane.

OpenSearch is not a panacea—while fuzziness can generate too many search results, it can also miss true positives where the OCR is particularly garbled. Fuzzy search parameters create a fundamental trade-off between over-inclusiveness and under-

[21]Replacements, insertions, deletions, adjacent transpositions. Simple "edit distance" or "Levenshtein distance" does not permit adjacent transpositions. For example, the following are all Damerau-Levenshtein edit distance of one; replacement: bat→bot, insertion: cat→cart, deletion: bread→read, adjacent transposition: form→from.

[22]*Recall* is the ratio of items correctly labeled positive by a model over the number of positives in the underlying dataset—(True Positives) / (True Positives + False Negatives) or the percent of relevant labels found. *Precision* is the ratio of items correctly labeled positive by a model over the total number of items labeled positive by the model—(True Positives) / (True Positives + False Positives) or the share of positive labels correctly given. These measures are particularly useful as they retain their meaning whether the positive class is very large or very small as a share of the total population, unlike simple accuracy. Simple accuracy is (correct predictions) / (total predictions). Imagine a dataset with ninety 1s and ten 0s—a model that always predicted 1 would have a simple accuracy of 90%, despite the fact that this is obviously a bad model for the purpose of identifying the 0s. In this case, the model would have a precision and recall of 0, correctly capturing the "always 1" model's poor performance.

[23]This is not an endorsement of OpenSearch in particular. Many other full search engines are available such as Elasticsearch, Lucene, Solr, and Tantivy.

[24]Full-text search greatly extends the concept of keyword or string searching. It breaks text down into tokens (akin to, but not exactly words), keeps track of the relationship of tokens to each other, removes punctuation and capitalization, and stems words (essentially finding the root of words; both "running" and "ran" are stemmed to "run"). Terms and their positions are then indexed in a way that allows for fuzzy word search, fuzzy phrase search (e.g., edit distance at the word level, extra words, differing order), and more. This process both increases search flexibility and performance.

[25]See https://github.com/nasa/cmr-opensearch.

inclusiveness, requiring iterative refinement of search strategies that balance precision (percentage of predicted positives that are true positives) against recall (percentage of all true positives identified). Errors are expected in any human produced text—spelling, grammatical, and typographical errors—these are generally small in edit distance and can be much improved via a simple spellchecker. However, our dataset has materially more issues due to image quality and underlying transcription method variation, greatly impacting OCR accuracy for a substantial fraction of our data. The wide range of possible errors and issues contained in the text extracted from the deed images means that we cannot achieve high recall and high precision simultaneously with fuzzy search alone.

By design, fuzzy search casts a broad net and returns inexact matches to the search term. There are many standard metrics used to measure the degree of fuzziness or how inexact matches can be. OpenSearch's most permissive available measure is DL distance with a cap of 2 edits.[26] Whether this is an appropriate restriction is highly dependent on the search term and the data quality. On one hand, it can result in too many matches—"interstate" is DL distance 1 from "intestate" and is a word we expect to see fairly often in contexts well outside of intestate property transfers. Thus, in practice, some fuzzy searches in OpenSearch for terms of interest return tens of thousands of matches of which only a handful are true positives.

On the other hand, fuzzy search can also yield too few matches due to OpenSearch's hard limit of DL distance 2.[27] For example, "diedintestate" will not match "intestate" as its DL distance is 4. Given OCR and transcription errors, it is not uncommon for terms of interest to sit well outside DL distance 2. To overcome this issue, we employ an expansive search strategy that begins with a comprehensive list of terms likely to capture relevant deeds, then iteratively refine this list by incorporating fuzzy search variants derived from verified positive matches. For example, once we know "interstate" can sometimes be a true positive in our data, we then use "interstate" as the search term in our query. The addition of the "r" effectively circumvents OpenSearch's DL distance limit and produces a query where the maximum DL distance from "intestate" is actually 3. Similarly, a fuzzy search using the term "diedintestate" mostly overcomes issues of dropped spaces and further increases the maximum DL distance compared with only searching for "intestate." While this technique does unearth previously unlocated true positives, it quickly moves us from the problem of too few matches to the opposite

[26] See footnote 21 for a detailed definition of DL distance. OpenSearch's hard cap of DL distance 2 means that "porches" will match "porch" (two deletions) but not "ponch" (two deletions plus one substitution = DL distance of 3). However, "porches" can easily become "ponch" through common OCR errors: missing the "es" ending and misreading "r" as "n". Porches are a commonly restricted feature in deed covenants.

[27] Even if we were able to search on fuzziness greater than DL 2, we would see many false positives for most terms of interest.

problem of too many matches, i.e., false positives.

Unfortunately, there does not exist any "one size fits all" rule that can perfectly mitigate either of the issues discussed above. Rules by themselves will yield suboptimal results compared with those achieved by ML or LLMs.

4.3 The Role of LLMs

As explained above, OCR and OpenSearch alone do not allow us to find a needle in the haystack but do reduce the haystack to a hay bale. This is where a custom model or LLMs come into play. Frontier LLM models do a good job comprehending simple concepts even within very garbled OCR text. However, it is quite expensive to use these models to label or search an entire corpus. For example, answering one simple question such as "does this deed contain a covenant restricting what type of business can be run on the property?" for all 4.7 million images in our sample would cost about $13,800 in input tokens alone and close to $24,000 in total.[28] If the prompt needed tweaking or we wished to ask about a different type of restriction, this cost would be incurred again. An alternative would be to pose multiple questions in a single prompt during one pass through the corpus, but this approach has been shown to meaningfully decrease the LLM's performance of the labeling task (OpenAI, 2025).

OpenSearch allows us not only to find phrases of interest but to pinpoint where in the document the phrase resides. Using this location, we can then extract a snippet of text around the phrase—providing textual context to the phrase of interest. We then subject this snippet to LLM interrogation. This is faster and much less costly than submitting text from an entire deed or image because increasing the relevance of the text submitted to the LLM decreases the false positives and time spent in extraneous human review. Since LLM performance degrades as a function of the length of input information, this targeted approach also *improves accuracy* (OpenAI, 2025). For example, in our work examining covenants restricting ownership, use, or occupancy, false positives were 2.4 times greater when considering a whole page vs. a snippet of interest. Evaluating the entire page produced only one additional true positive, or a 0.01% increase, which was not located when using only snippets.[29]

Figures 4, 5, and 6 present several examples of an LLM in action differentiating true positive and false positive matches for covenants restricting permissible commercial

[28] LLMs charge separately for input and output tokens—i.e., for the input data and question(s) posed to the LLM and the output response from the LLM.

[29] These numbers quantify the performance of the LLMs labeling accuracy only. Error in highlighting the right snippet on a page of text increases this number to 0.5% more true positives when using the whole page. Above, we exclude these cases to focus only on LLM performance on longer input vs. combining this with how well we extract passages of interest with OpenSearch.

uses of property. These examples show how we ask an LLM[30] to review a candidate match found by an OpenSearch query. We provide an initial "system prompt," which provides the primary framework that determines how the LLM will interact throughout the remainder of the "conversation."[31] In testing,[32] a sizable challenge was ensuring the LLM's response was consistently formatted[33] in such a way that we could easily automate large numbers of prompts[34] and responses. LLMs in general have been trained to include their "thought process" in their response to better facilitate validation of their response by the user. While well-intentioned, this "thought process" was the largest hurdle to producing consistently formatted responses. We found that the key to overcoming this hurdle was to provide a specific location in our desired response format[35] in which the LLM could store its "thought process."

Below, we show the system prompt (Fig. 4) and two prompt and response interactions using actual data—one each for true (Fig. 5) and false positives (Fig. 6) for commercial use restriction. Human submitted prompts to the LLM are shown with the header "Human" and an icon showing a person. Example responses from an LLM are shown with the header "Assistant" and a robot icon. These visuals reflect the typical interface for interacting with LLMs and demonstrate what our workflow looks like in practice.

[30]Claude 3.7 Sonnet for these examples. This is not an endorsement of Claude in particular.

[31]This is where you can instruct an LLM to respond like Shakespeare or at a level that a child could understand.

[32]Referred to as "prompt engineering."

[33]We chose to use standard JSON format to structure the LLM response. JSON stores data values as key-value pairs, like a dictionary—"LLM": "large language model."

[34]"Prompt" is the term used for any action a human takes in the LLM interface, i.e., asking a question, providing text to summarize, or asking the LLM to perform a task.

[35]We called this location in our format "comments."

Figure 4: The system prompt provided to the LLM instructing it how to interact with the data and how to format its response. The "comments" section provides a space for the LLM to explain its reasoning.

Human:

You are a lawyer very carefully reading deeds to find covenants restricting use of property for business purposes. It is important that you read documents very carefully and in their entirety. The text you are reading will be OCR'd text and will have some mistakes, you should try to correct these. You must very carefully read and consider the document, taking in to account the full context of the text I present. First consider how you would respond in a more lengthy discussion where you explain things, then summarize this into just a yes or no answer. You response will be in the form of the following json, where you use proper escaping for quotes within strings using backslashes:

```
{
    "covenant_exists": boolean
    , "comments": string
}
```

Absolutely do not return anything but json. If you have more to say put it in "comments". Do not use any quotation marks of any form inside of the body of "comments". Only use them as appropriate to denote a string in a json.

Figure 5: A prompt and response example for a text snippet containing a true positive for a business restriction. The snippet is correctly identified as a true positive by the LLM.

Human:

Does this snippet of a deed contain a covenant restricting what types of business can be run on the property?

UNDER AND SUBJECT to the expyess restriction and condition that no gasoline station eervice station automobile accessory shop delicsteefen store or drug stove shall be erected upon any part of the h ereby grant ed premises neither shall the building er ect ed thereon be used for any of the above mentioned purposes noy shall any building thereon erect ed *be used as a store OT be used roy any offensive use or occupation within Fifteen years fyom the date hereof nor shall any building crect ed th ey eon or that may be erect ed thereon be ext ended any nearer the building line of Martine Mill Road than the now is within Firteen years from the dat e hereof.

Assistant:

{
    “covenant_exists”: true,
    “comments”: “The deed contains explicit business restrictions prohibiting gasoline stations, service stations, automobile accessory shops, delicatessen stores, and drug stores. It also includes a broader restriction that no building shall be used as any type of store or for any offensive use or occupation within fifteen years from the date of the deed. The covenant contains OCR errors but its business restrictions are clear.”
}

Figure 6: A prompt and response example for a text snippet identified to contain a phrase of interest that does not actually contain a business restriction (a true negative). Leases and other agreements are recorded with deeds in Philadelphia. This snippet belongs to the lease of a drugstore rather than a property deed restricting the use of the building. However, the similarity in language leads to the LLM classifying the snippet as containing a restriction, an example of a false positive.

Human:

Does this snippet of a deed contain a covenant restricting what types of business can be run on the property?

farther consideration for the said hiring and letting to the said pasty of the second part that he the said party of the second part thall conduct seid Drag Stere in his name and that he will punctually pay the hire as aforesaid on the days and times above pextion mentioned and that he will take good care of the said Drug Store and will and Conduct of the sald business up totte standard

Assistant:

{
“covenant_exists”: true,
“comments”: “This deed snippet appears to be about leasing or hiring a Drug Store business rather than a property covenant. It contains terms about how the party of the second part shall conduct the Drug Store business (in his name, paying punctually, taking good care of it, conducting it up to standard), but does not restrict what types of businesses can operate on the property. The text describes operational requirements for a specific business arrangement rather than a land use restriction. There are several OCR errors in the text (thall should be shall, seid/sald should be said, totte should be to the).”
}

In our work examining covenants restricting ownership, use, or occupancy, the performance of the LLM is quite good. Approximately 98% of the text snippets the LLM labels as positives are correct (precision), and it locates between 90-98% of all positives (recall). That said, it is essential that we do not take the LLM result as the final answer. To facilitate validation, we developed a tool that enables humans to review labels generated by the LLM. The tool presents the human reviewer with text snippets containing the terms matched in OpenSearch and the LLMs label and reasoning. A human reviewer can then quickly validate the LLMs labels, referring to the source image where necessary.

The combination of OCR, full-text search, and LLMs allows us to drill down to a focused set of deeds for which a human can verify and quantify the LLM's performance. We send approximately 130,000 high- to medium-propensity[36] OpenSearch matches through an LLM. We then manually review on the order of 30,000 of the resulting LLM labels. This subset is composed of all LLM positive labels and a sample of LLM negative labels for snippets we most expect to be true positives, given the OpenSearch query terms. We seek to fully quantify precision and recall, which requires substantial verification of LLM-produced negative labels. At a minimum, all positive LLM labels and all negative LLM labels of very high-propensity OpenSearch matches should be reviewed. For example, when searching for intestate transfers, if OpenSearch matches "intestate" *exactly* with no edits and an LLM labels this transaction as not an intestate transfer, then this LLM negative label should be manually verified. In our case, the minimum viable number of positive and negative LLM labels for review totaled less than 9,000 snippets. We review substantially more to achieve more accurate estimates of false negative and recall rates.

We also used LLMs to evaluate a substantial sample of low to very low propensity OpenSearch matches—100,000 images out of 4.5 million. Among this sample of low-probability images, we reviewed all 111 positive LLM labels and found only 17 true positives. This high false positive rate was expected, given the very low probability that these images would contain relevant content.

Overall, we found that only 0.016% of the sample images contained true positives. Due to cost constraints, we did not apply LLMs to all remaining text. However, if we ran the remainder of our dataset through an LLM, we estimate we could find an additional 408[37] deeds of interest—a 5.5% increase over our current total (Fig. 7). This comprehensive analysis would cost approximately $22,262, or $55 per additional true positive found. We could reduce costs by focusing only on deeds recorded before 1950, after which the language of interest becomes extremely rare. This targeted approach would cost about $36 per true positive and yield approximately 380[38] additional matches (94% of the total possible additional discoveries for 62% of the cost). However, finding these matches would require reviewing approximately 2,500 text snippets, 85% of which would be false positives. In contrast, we found our first 97% of true positive matches,

[36]By "propensity," we mean the approximate probability of a true positive, given the OpenSearch terms used. For example, when searching for intestate transfers, an exact match or edit distance 1 to "intestate" would be a high-propensity match.

[37]We are certainly missing less than 5.5% of properties associated with our restrictions of interest. Properties are likely to transact multiple times over our 56-year window—we'd need to miss them multiple times to not have knowledge of their treatment state.

[38]It is worth noting we would still not have found all possible deeds of interest. In cases where no legible indicative language was found via OpenSearch, LLM recall was around 60-80%. Assuming omnipotence, this would boost the 408 to 510-680 and 380 to 475-633, perhaps as much as 9% of the universe we have identified.

or 91% factoring in the estimated 408 that we have not found due to cost constraints, for just $0.05 per true positive.

Fig. 7 shows that the incremental cost of finding additional true positives increases as a function of how likely the search terms used in OpenSearch are to identify documents containing true positives. As you move along the x-axis, the corpus of documents being reviewed contains less specific language as identified via OpenSearch because it did not match high-propensity words or phrases. For example, starting with near exact matches to highly specific legal language, moving to less precise language with more fuzziness, and finally to the entire remaining corpus. The gains to labeling unlikely documents using an LLM are low. Our adaptive modular use of OpenSearch allows us to spend LLM tokens only where they are likely to achieve an efficient return. Fig. 8 is a cumulative version of the prior figure, showing the point of extreme diminishing return.

While we didn't strictly need to expand the LLM application beyond high-probability candidates, this exercise helped us quantify the benefits and limitations of our adaptive modular approach combining OpenSearch and LLMs. By using OpenSearch to pre-screen for high-probability candidates before applying LLMs, we reduce false positive review costs and significantly reduce true positive detection costs while maintaining high recall rates.

Figure 7: The percentage of LLM-identified true positives (TP) compared with the average cost per each LLM-identified TP. The incremental percentage is scaled by the total TP correctly labeled by the LLM plus the estimated number of remaining true positives that our LLM would identify if run on the entire corpus. The cost per TP increases as the search criteria used in OpenSearch become less restrictive, fuzzier, and less likely to identify documents containing true positives.

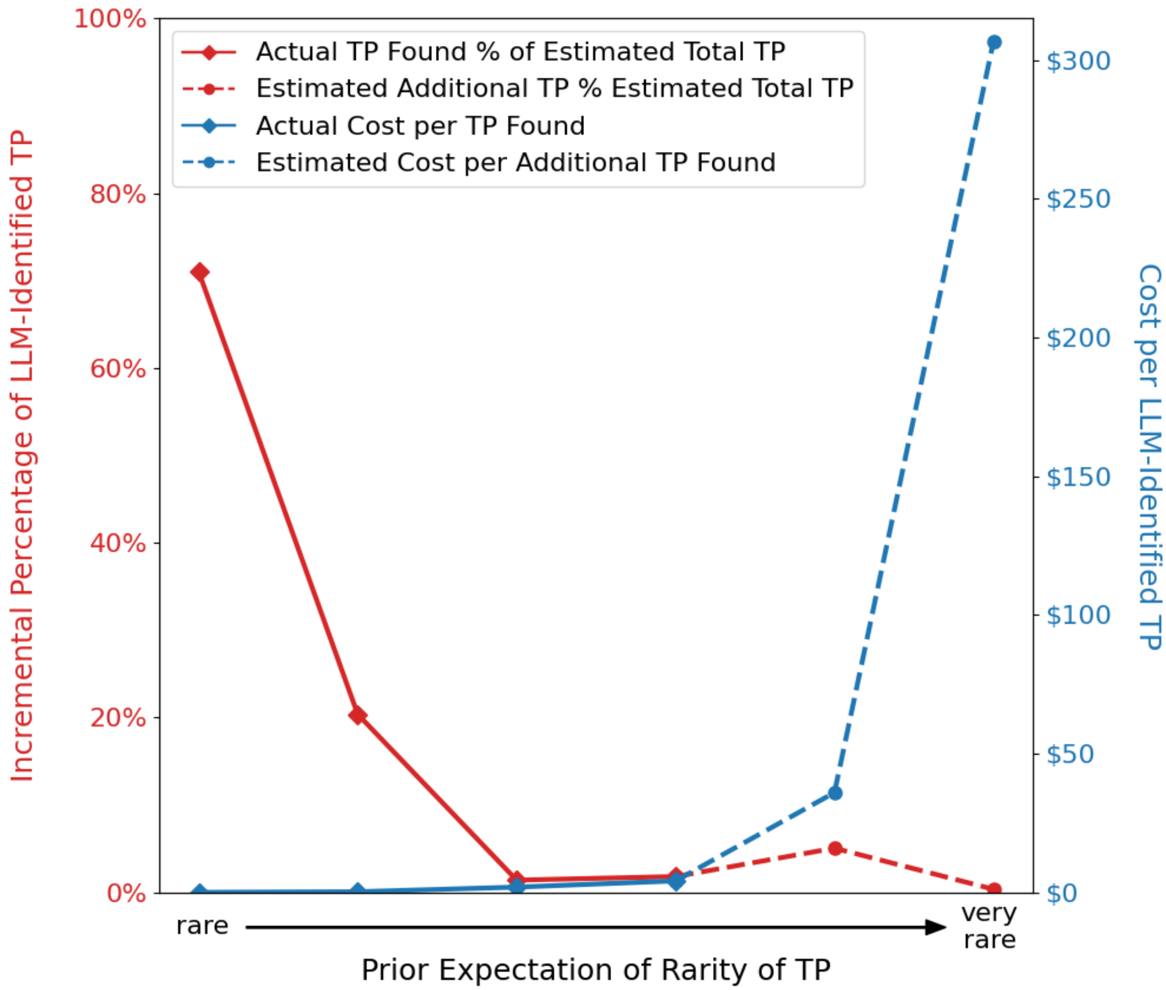
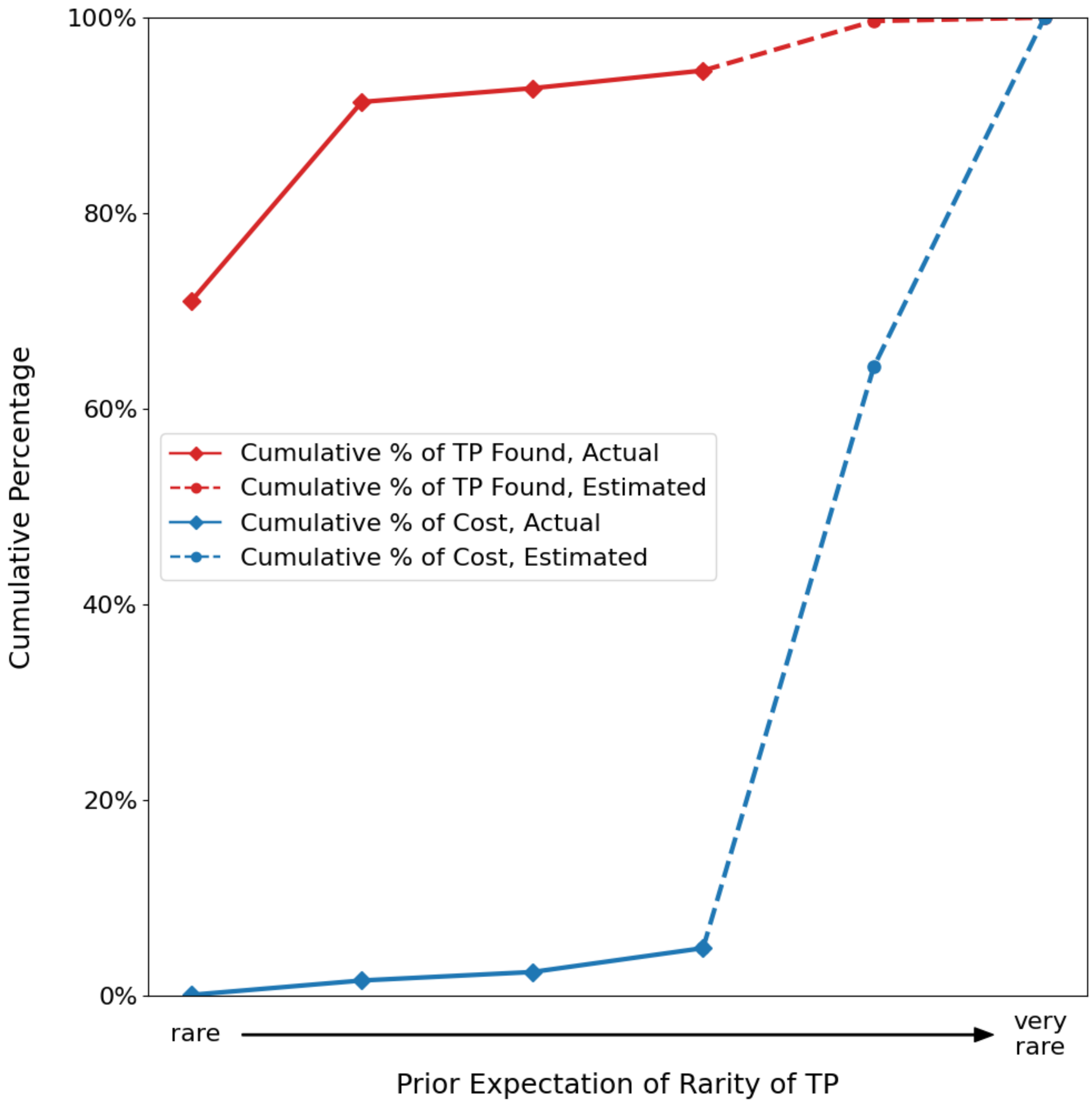


Figure 8: Cumulative percentage of LLM-identified true positives compared with cumulative percentage of cost. Cumulative percentage TP is scaled by the total estimated number of TP. The bulk of a project's cost is incurred identifying the "last few" true positives.



4.4 The Final Step: Geolocation

The final challenge is to geolocate the deeds of interest. Geolocation allows us to quantify the localized and neighborhood effects of deed terms and restrictions on the property market.[39] Geolocation also helps to link property transfers through time.[40] Since it is common for restrictions to be enumerated in detail only in the deed in which they are first put in force, tracking properties through time is necessary to identify the specific terms applying to a given property. Exact identification of deed terms, in addition to when and where these terms apply, makes geolocation an essential step for econometric analysis of deeds.

Address-based geocoding would be challenging but feasible if the deeds consistently contained the address of the property being transacted. In such cases, we could isolate the address string with an ML model or LLM and attempt to geolocate it—somewhat difficult considering that historical addresses may no longer correspond to current street names or numbering systems. However, this method provides limited usefulness, as only about 30% of deeds prior to 1935 contain addresses.

Fortunately, Philadelphia deeds always contain a "metes and bounds" description of the property. This is a textual, legal description of the location and boundaries of a property as determined by a surveyor relative to a "stable" point of reference.[41] This description can be converted into geographic coordinates and a corresponding parcel polygon (Wex Legal Dictionary, 2025). Depending on factors like the shape of the parcel and the legibility of the deed image, the degree of difficulty in converting metes and bounds descriptions to geographic coordinates ranges from moderate to extreme.

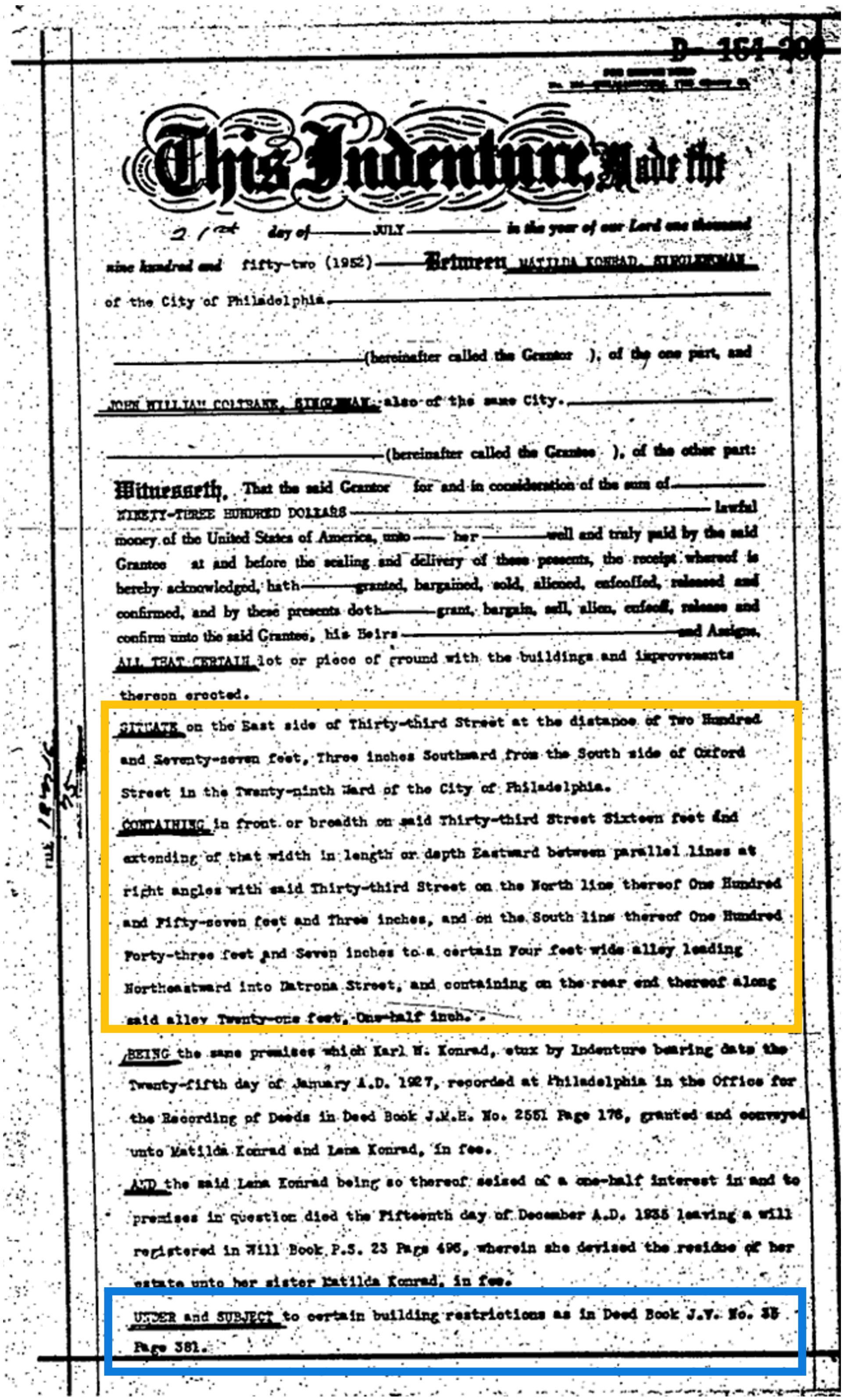
The case of John Coltrane's house, a National Historic Landmark in the Strawberry Mansion section of Philadelphia, highlights some of the very issues mentioned earlier, including a tangled title that has stifled preservation efforts for many years (Jackson, 2024). We use this deed as an example of translating a relatively straightforward metes and bounds description into a geolocated parcel. Figure 9 presents the first page of the deed recording Coltrane's purchase of the home. The blue box indicates that the property is subject to building restrictions not described in this deed. This serves as an example of the importance of the second use case of geolocation described above—linking properties through time with the assistance of physical location. The metes and bounds property description is highlighted by the yellow box.

[39] For example, covenants restricting ownership, use, or occupancy or transfer under intestacy or by will etc.

[40] Geolocation can be used in conjunction with buyers, sellers, and references to previous property transfer information, such as dates and the book and page in which the prior deed was recorded.

[41] At times, the point of reference can be something surprisingly fuzzy, such as "the large elm on the northeast corner of the Smith farm."

Figure 9: First page of the deed describing John Coltrane's purchase of his Philadelphia home. The yellow box outlines the metes and bounds description of the property. The blue box highlights the section of the deed listing the property restrictions. In this case, the property restrictions refer to a previous deed rather than stating the restrictions in full.



D-164-200

This Indenture, Made the 21st day of JULY in the year of our Lord one thousand nine hundred and fifty-two (1952) Between MATILDA KONRAD, SINGLEWOMAN, of the City of Philadelphia (hereinafter called the Grantor), of the one part, and JOHN WILLIAM COLTRANE, SINGLEMAN, also of the same City. (hereinafter called the Grantee), of the other part:

Witnesseth, That the said Grantor for and in consideration of the sum of NINETY-THREE HUNDRED DOLLARS lawful money of the United States of America, unto her well and truly paid by the said Grantee at and before the sealing and delivery of these presents, the receipt whereof is hereby acknowledged, hath granted, bargained, sold, aliened, enfeoffed, released and confirmed, and by these presents doth grant, bargain, sell, alien, enfeoff, release and confirm unto the said Grantee, his Heirs and Assigns,

ALL THAT CERTAIN lot or piece of ground with the buildings and improvements thereon erected.

SITUATE on the East side of Thirty-third Street at the distance of Two Hundred and Seventy-seven feet, Three inches Southward from the South side of Oxford Street in the Twenty-ninth Ward of the City of Philadelphia.

CONTAINING in front or breadth on said Thirty-third Street Sixteen feet and extending of that width in length or depth Eastward between parallel lines at right angles with said Thirty-third Street on the North line thereof One Hundred and Fifty-seven feet and Three inches, and on the South line thereof One Hundred Forty-three feet and Seven inches to a certain Four feet wide alley leading Northeastward into Natrona Street, and containing on the rear end thereof along said alley Twenty-one feet, One-half inch.

BEING the same premises which Karl W. Konrad, etux by Indenture bearing date the Twenty-fifth day of January A.D. 1927, recorded at Philadelphia in the Office for the Recording of Deeds in Deed Book J.M.H. No. 2551 Page 176, granted and conveyed unto Matilda Konrad and Lena Konrad, in fee.

AND the said Lena Konrad being so thereof seized of a one-half interest in and to premises in question died the Fifteenth day of December A.D. 1935 leaving a will registered in Will Book P.S. 23 Page 495, wherein she devised the residue of her estate unto her sister Matilda Konrad, in fee.

UNDER and SUBJECT to certain building restrictions as in Deed Book J.V. No. 35 Page 381.

Figure 10 illustrates how the metes and bounds description of the Coltrane house translates to a description of the geographic location and shape of the associated parcel. The right half of Figure 10 provides a map of Coltrane's property; the blue region denotes the exact parcel. The left half of Figure 10 provides an exact[42] transcription of the metes and bounds text highlighted in the yellow box in Figure 9. The color of the text in the metes and bounds description matches the lines on the map corresponding to the steps taken to trace the outline of the property's boundary, starting from a point of reference and ending in a closed polygon representing the parcel.

Figure 10: Transcribed text of the metes and bounds description of John Coltrane's home. This text is an exact transcription without OCR errors. The blue shaded region is the actual parcel of the property. The color of the text has been matched to the color of the corresponding line described by the text.

SITUATE on the East side of Thirty-third Street at the distance of Two Hundred and Seventy-seven feet, Three inches Southward from the South side of Oxford Street in the Twenty-ninth Ward of the City of Philadelphia.

CONTAINING in front or breadth on said Thirty-third Street Sixteen feet and extending of that width in length or depth Eastward between parallel lines at right angles with said Thirty-third Street on the North line thereof One Hundred and Fifty-seven feet and Three inches, and on the South line thereof One Hundred Forty-three feet and Seven inches to a certain Four feet wide alley leading Northeastward into Natrona Street, and containing on the rear end thereof along said alley Twenty-one feet, One-half inch.

This example uses an exact transcription of a fairly simple property description and shape, but this is not at all typical. OCR often introduces text errors that, when combined with highly complicated descriptions of property boundary lines[43] and/or ambiguous starting points,[44] make georeferencing a challenge for humans and machine alike. In our application, neither direct elicitation[45] of address nor translation of metes

[42] This example contains no OCR errors; it is an exact transcription.

[43] Metes and bounds descriptions can be extremely long—from several paragraphs to several pages—and describe complicated polygons with dozens of segments both straight and curved.

[44] A reference or starting point can be described as something like "1017 feet 8-7/8 inches Northwestwardly from an ***old stone*** at the intersection of the middle of School House Lane (40 feet wide) and the middle of said Wissahickon Avenue." This is actual text that is part of the metes and bounds description from one of our deeds of interest from 1925.

[45] As a large fraction of our deeds do not contain an explicit address.

and bounds are particularly useful for large-scale automated georeferencing.

To address these limitations, we are developing an alternative approach focused on achieving block-level or street-corner precision rather than exact parcel identification. Our approach relies on the structural logic of metes and bounds descriptions within a city street grid. This logic is fairly consistent across any East Coast city laid out in a grid.[46] Each description must reference at least two intersecting streets. The first street mentioned typically indicates the property's frontage, while the intersection of the first and second street serves as an anchor for the rest of the boundary description. Additional street references (e.g., a third or fourth street) will describe the rest of the closed city block containing the property. The approach leverages the fact that there is a finite set of valid reference street combinations that define the boundaries of any closed city block. By analyzing the frequency and sequence of street name mentions against the set of valid street combinations, we score matches to probabilistically assign deeds to their most likely geographic locations. This allows, in many cases, not just block-level georeferencing but block corner precision, i.e., saying that the property is inside the city block formed by streets a, b, c, d and is closest to the corner of a and b. This method will allow us to geolocate a substantial portion of properties—at least two-thirds of them. We are developing additional geolocation techniques for rooftop precision. Regardless, our block-level geolocation makes human translation of metes and bounds to an exact parcel shape easier and faster.

This geolocation method takes advantage of our adaptive modular approach. We adopted full-text search with the primary purpose of identifying language of interest. Given the flexibility of this tool, we are also able to use it to perform the georeferencing logic mentioned above. With OpenSearch, we can search documents for combinations of street names corresponding to real block combinations—identified via street data sourced from the City of Philadelphia. OpenSearch also allows us to understand where in the document and in what order these streets are mentioned. This allows us to identify the most likely block corner matching a particular deed. Thus, with no incremental data processing or infrastructure work, we are able to implement logic for georeferencing deeds.

[46] Material variations arise in Western cities where land is delineated with reference to the Public Land Survey System – a universal, regular grid. Older US cities use metes and bounds, mostly inherited from England, based on reference to key natural or manmade features. In cities, this mostly means street intersections.

5 Deeds: Next Steps

We are now diving deeper into a subset of deeds of interest identified via our adaptive modular approach. For approximately 7,500 deeds,[47] we are working with a data labeling firm to gather high-quality labeled data—that is, to exactly transcribe text of interest, and to associate this text with its exact location on the page, i.e., enumerating the "bounding box" containing the text. Labeled data themselves are valuable—they will allow econometric measurement of the effects of specific types of deed restrictions. They will also allow us to make an informed decision on whether further hyperspecialized modeling will yield worthwhile improvements or new insight.

Our restrictive language of interest is very rare as a percentage of all deeds recorded between 1910 and 1965—less than one-quarter of a percent.[48] In ML terms, we would say that there is severe "class imbalance"—negative cases are orders of magnitude more common than positive cases. In more common parlance, one might say that the signal-to-noise is small. Any standard sampling strategy would have resulted in far too few (likely no) positive cases in the training data and therefore not enough positive cases to successfully train a model. Class imbalance is frequently addressed by oversampling the rare class—which can be very difficult when you don't yet know which cases are the rare class, i.e., in completely unlabeled data. Our adaptive modular approach allows us to focus our work on a sample already selected for syntactical context—i.e., words often or always associated with our restrictive language of interest. This highlights why we believe our adaptive modular approach is the most efficient approach to working with historic data, especially in cases with low signal-to-noise.

6 Conclusion

Historical administrative text data, which has been largely inaccessible for analysis due to format constraints, can now be efficiently processed and analyzed at scale, thanks to recent technological advances. At the Center for the REstoration of Economic Data (CREED), we are working to understand, harness, refine, and customize these advancements to produce high-quality publicly available data, code, and economic research. In this paper, we evaluate two competing paradigms for large-scale historical document digitization and demonstrate why adaptive modular methodological approaches can offer better performance across the entire research process.

[47] Each deed on average covers multiple properties. We are working on an accurate estimate of property count.

[48] They are significantly less rare as a percentage of all properties in Philadelphia, as deeds on average cover multiple properties.

References

Aranjuelo, N., García, S., Loyo, E., Unzueta, L., & Otaegui, O. (2021, June). Key Strategies for Synthetic Data Generation for Training Intelligent Systems Based on People Detection from Omnidirectional Cameras. *Computers & Electrical Engineering*, *92*(107105), 107105. Retrieved from https://doi.org/10.1016/j.compeleceng.2021.107105

Asquith, B. J., Mast, E., & Reed, D. (2023, March). Local Effects of Large New Apartment Buildings in Low-Income Areas. *Review of Economics and Statistics*, *105*(2), 359–375. Retrieved from https://doi.org/10.1162/rest_a_01055

Bartik, A., Gupta, A., & Milo, D. (2025, August). The Costs of Housing Regulation: Evidence From Generative Regulatory Measurement. *SSRN Electronic Journal*. Retrieved from https://dx.doi.org/10.2139/ssrn.4627587

Bauer, A., Trapp, S., Stenger, M., Leppich, R., Kounev, S., Leznik, M., ... Foster, I. (2024, February). *Comprehensive Exploration of Synthetic Data Generation: A Survey.* Retrieved from https://arxiv.org/abs/2401.02524

Hincken, G. (2021, August). *How Tangled Titles Affect Philadelphia* (Report). The Pew Charitable Trusts. Retrieved from https://www.pew.org/en/research-and-analysis/reports/2021/08/how-tangled-titles-affect-philadelphia

Hughes, C., Jr. (1958, April). The Philadelphia Program. *American Archivist*, *21*(2), 131–142.

Jackson, M. (2024, June 21). A New Initiative to Protect Black History Starts With Coltrane. *The New York Times*. Retrieved from https://www.nytimes.com/2024/03/05/realestate/john-coltrane-house-philadelphia-black-history.html

Korngold, G. (2001). The Emergence of Private Land Use Controls in Large-Scale Subdivisions: The Companion Story to *Village of Euclid v. Ambler Realty Co.* *Case Western Reserve Law Review*, *51*(4), 617–643. Retrieved from https://scholarlycommons.law.case.edu/caselrev/vol51/iss4/5

Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023, December). Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 10443–10461). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.647/ doi: 10.18653/v1/2023.emnlp-main.647

Liu, R., Wei, J., Liu, F., Si, C., Zhang, Y., Rao, J., ... Dai, A. M. (2024, August). *Best Practices and Lessons Learned on Synthetic Data.* Retrieved from https://arxiv.org/abs/2404.07503

Meng, X.-L. (2023, April). Data Science and Engineering with Human in the Loop, Behind the Loop, and Above the Loop. *Harvard Data Science Review*, *5*(2).

OpenAI. (2025). *Context Arena.* Retrieved from https://contextarena.ai/ (Accessed on July 8, 2025.)

Santucci, L. (2020, November). Documenting Racially Restrictive Covenants in 20th Century Philadelphia. *Cityscape: A Journal of Policy Development and Research*, *22*(3), 241–267.

Stein, S., & Carpenter, A. (2022, January). *Heir's Property in an Urban Context* (2022 AEA/ASSA Annual Meeting VIRTUAL, January 7-9, 2022). Agricultural and Applied Economics Association. Retrieved from https://doi.org/10.22004/ag.econ.316788 doi: 10.22004/ag.econ.316788

Storey, E. A. (2018, April). Cartographically Reconstructing Surveys of Community Land Grants in New Mexico to Support Historical Research and Political Discourse. *Historical Methods*, *52*(2), 95–109.

Surani, F., Suzgun, M., Raman, V., Manning, C. D., Henderson, P., & Ho, D. E. (2025). *AI for Scaling Legal Reform: Mapping and Redacting Racial Covenants in Santa Clara County.* Retrieved from https://arxiv.org/abs/2503.03888

Weiss, M. A. (1987). *The Rise of the Community Builders.* Frederick, MD: Beard Books.

Wex Legal Dictionary. (2025). *Metes and Bounds.* Retrieved from https://www.law.cornell.edu/wex/metes_and_bounds (Accessed on August 27, 2025. Updated in August 2020.)