An Evaluation of GPT-4V for Transcribing the Urban Renewal Hand-Written Collection

Lee, Myeong
mlee89@gmu.edu
George Mason University, Fairfax, U.S.

Hsu, Julia H. P.
hhsu2@gmu.edu
George Mason University, Fairfax, U.S.

1   Introduction

From the early 1960s to the mid-1970s, Asheville, North Carolina, underwent urban renewal, a national program aimed at modernizing "blighted" areas (Lee et al. 2017). This process, mostly impacting African-American neighborhoods, displaced families, businesses, and organizations for economic and infrastructure development. Due to its historical significance and on-going reparation efforts in many cities in the United States, understanding parcel-level information about how each property was acquired by the housing authority is one of the most important steps. This paper focuses on the East Riverside neighborhood (now Southside), the southeastern U.S.'s largest area affected by urban renewal. Our target documents are property acquisition documents originally produced by the Housing Authority of the City of Asheville (HACA) that details the property acquisition processes of nearly 1,000 properties in Southside. They include property's complete sales history, including appraisals, offers, rejections, and court cases.

However, the extensive lengths (20-200 pages per parcel) and complexity of property acquisition documents from the urban renewal era pose significant challenges in digital curation, particularly for handwritten documents. This is a persistent issue for computational archival scientists and digital humanities scholars. Although handwritten text recognition (HTR) platforms such as Transkribus (Kahle et al. 2017) and Google Document AI[1] have been extensively studied with good performances (e.g., (Ingle et al. 2019, Kiessling et al. 2019)), their practical application by non-experts remains challenging. This is due to the additional technical efforts required to configure recognition systems, such as training for specific document types, data structuring, and cleaning processes.

In November 2023, OpenAI released GPT-4V(ision), which includes Optical Character Recognition (OCR) capabilities. Given that much of the data curation, processing, and cleaning can be managed through user-friendly prompts (i.e., chat), we aim to conduct an initial assessment of GPT-4V's effectiveness in transcribing hand-written documents from the urban renewal collection. If GPT-4V can accurately digitize hand-written documents through carefully crafted prompts, it could become a valuable tool for non-experts in transcribing historical documents on a large scale. Alternatively, if it falls short, it is still crucial to understand and discuss the implications of using Large Language Models (LLMs) for digitizing archival documents.

This paper evaluates GPT-4V's performance in transcribing the cover pages of selected urban renewal documents. These cover pages, all hand-written and generally more challenging to read (even for humans) compared to other parts of the documents, are valuable for researchers and practitioners focusing on urban renewal, as they succinctly provide key information about property acquisition processes. Figure 1 shows an example of an original document.

2   Approach

We utilized the OpenAI APIs [2] to make queries to GPT-4V. We employed constrained prompting techniques to instruct GPT-4V to read the text in the document (Yang et al. 2023). The objective was to retrieve information from individual fields and organize it into a specific JSON format. We engineered the prompt following the guideline by Yang et al. (2023). Through iterative refinements of the prompts, the constrained prompt was finalized as
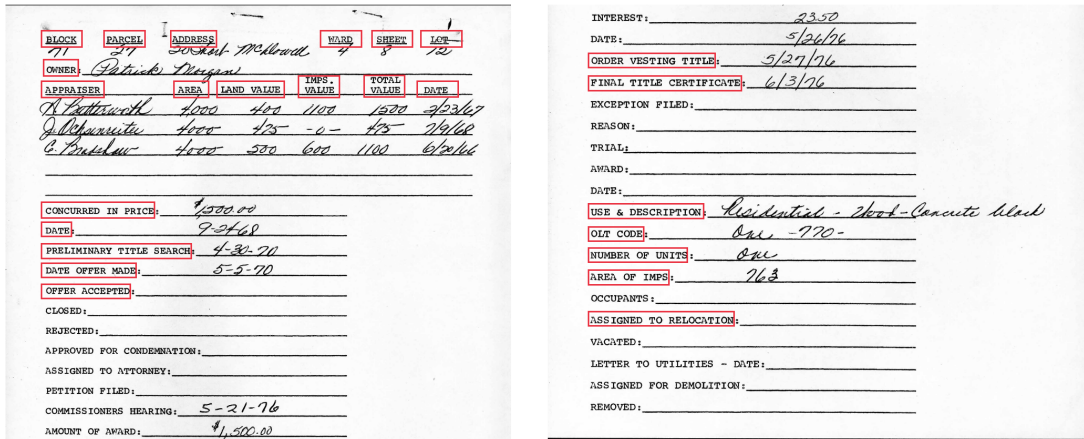
Figure 1: An example of original document. Red rectangles are fields of our focus in transcription.

depicted in Figure 2. Using this prompt, we processed 50 select documents, subsequently assessing its performance per each field.

This process involved comparing the text within the retrieved JSON data by GPT-4V with the ground truth data compiled by human evaluators. Given that several fields were frequently empty across most documents (e.g., Exception filed, Trial), our focus was on fields with values present in more than 50% of the samples (25 documents). It is worth noting that the accuracy was calculated solely based on non-empty values. Based on the accuracy of each field's transcribed values, we qualitatively examined the characteristics of the hand-written letters to understand the strengths and weaknesses of GPT-4V.

3 Results

The processing time for GPT-4V to read 50 documents was about 19.11 minutes. Table 1 presents the accuracy for each individual field. The *Accuracy* column of Table 1 reports the accuracy solely based on non-empty values. Notably, the field *ward* achieved the highest accuracy score of 0.8, whereas *owner* obtained the lowest accuracy score of 0.18 in this evaluation. Our initial assessment suggests that, without further engineering of prompts and additional frameworks that help refine the results, GPT-4V is not at the stage of generalizable transcription tasks for hand-written texts.

To understand the reasons for the varying accuracy in transcribing tasks, we qualitatively analyzed the hand-written texts by comparing between different fields. Notably, fields that present high accuracy such as *ward* and *block* are more likely to provide numerical values in relatively simple forms, providing clear structures for comprehension by GPT-4V. Among the fields with more than 0.7 of accuracy are all simple numeric fields or dates, with an exception of the *use_description* field. Although the *use_description* field is English words, it presents relatively standardized terms, such as "Residential" or "Commercial", which may made it easy for GPT-4V to predict the values.

```
Please read the text in this image and return the information in
the following JSON format (note xxx is placeholder, if the
information is not available in the image, put 'N/A' instead). If
the appraisal is reviewed by addition person, put the information
in the first empty appraiser field. Ignore strikethrough text.
{
    "block": xxx,
    "parcel": xxx,
    "address": xxx,
    "ward": xxx,
    "sheet": xxx,
    "lot": xxx,
    "owner": xxx,
    "appraiser_1": {
        "name": xxx,
        "area": xxx,
        "land_value": xxx,
        "imps_value": xxx,
        "total_value": xxx,
        "date": xxx
    },
    ...
    "assigned_to_relocation  ": xxx,
    "vacated": xxx,
    "letter_to_utilities_date": xxx,
    "assigned_for_demolition": xxx,
    "removed": xxx
}
```

Figure 2: Constrained Prompting

Conversely, fields with low accuracy, such as *owner* and *appraiser_1_name*, are attributed to their longer, more diverse text entries, making it challenging for GPT-4V to extract specific information accurately. Finally, we identified consistent misreads of GPT-4V. They include, but are not limited to, a confusion between ”2” and ”7,” ”1” and ”7,” ”0” and ”6,” and ”*F*” and ”*E*”. These recurring errors illustrate consistent challenges faced by the model in distinguishing specific numerical and alphabetical characters accurately.

4   Discussion

Our findings offer valuable insights for using LLMs like GPT-4V to transcribe handwritten documents, particularly historical ones. Because LLMs exhibit varied performance in different types of texts, further examination is necessary for their application in transcribing handwritten texts. However, they may be suitable for simple tasks, as they tend to better recognize numerical and basic textual content.

Our research suggests several potential directions for future research. First, since GPT-4V commonly misreads certain characters (e.g., confusing 1 and 7), these patterns could be integrated into prompt design in a recursive manner. Second, for text-based fields, options could be recursively narrowed down through trial and error, potentially incorporating human oversight. Lastly, considering the diverse characteristics and performances of various LLMs across different handwriting styles, an ensemble approach combining multiple LLM tools could be developed to enhance transcription accuracy.

Acknowledgments

References

Ingle, R. R., Fujii, Y., Deselaers, T., Baccash, J. & Popat, A. C. (2019), A scalable handwritten text recognition system, *in* '2019 International Conference on Document Analysis and Recognition (ICDAR)', IEEE, pp. 17–24.

Kahle, P., Colutto, S., Hackl, G. & Mühlberger, G. (2017), Transkribus-a service platform for transcription, recognition and retrieval of historical documents, *in* '2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)', Vol. 4, IEEE, pp. 19–24.

Kiessling, B., Tissot, R., Stokes, P. & Ezra, D. S. B. (2019), escriptorium: an open source platform for historical document analysis, *in* '2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)', Vol. 2, IEEE, pp. 19–19.

Lee, M., Zhang, Y., Chen, S., Spencer, E., Cruz, J. D., Hong, H. & Marciano, R. (2017), Heuristics for assessing computational archival science (cas) research: The case of the human face of big data project, *in* '2017 IEEE International Conference on Big Data (Big Data)', IEEE, pp. 2262–2270.

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z. & Wang, L. (2023), 'The dawn of lmms: Preliminary explorations with gpt-4v (ision)', *arXiv preprint arXiv:2309.17421* **9**, 1.

Table 1: Accuracy of each field in the cover page of the urban renewal collection.

Field	Accuracy	# empty cells
ward	0.80	0
block	0.76	0
appraiser_1_land_value	0.74	0
order_vesting_title	0.74	8
use_description	0.74	8
sheet	0.72	0
appraiser_1_total_value	0.72	0
date_offer_made	0.69	8
offer_accepted	0.68	25
final_title_certificate	0.67	8
appraiser_2_total_value	0.63	1
preliminary_title_search	0.63	4
parcel	0.62	0
appraiser_2_land_value	0.61	1
appraiser_1_imps_value	0.60	25
appraiser_1_date	0.59	1
concurred_in_price_date	0.59	6
assigned_to_relocation	0.58	14
appraiser_3_date	0.57	8
appraiser_3_total_value	0.57	6
lot	0.56	0
concurred_in_price	0.54	2
appraiser_1_area	0.54	0
appraiser_2_date	0.53	1
appraiser_2_area	0.51	1
appraiser_3_name	0.41	6
number_of_units	0.41	23
appraiser_2_name	0.39	1
olt_code	0.31	15
address	0.30	0
area_of_imps	0.29	22
appraiser_1_name	0.22	0
owner	0.18	0