Enhancing Layout Understanding via Human-in-the-Loop: A User Study on PDF-to-HTML Conversion for Long Documents

Chenyu Mao

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Edward A. Fox, Chair
Sang Wong Lee
Yan Chen

Feb. 7, 2025
Blacksburg, Virginia

Keywords: ETD, deep learning, object detection, document layout analysis

Enhancing Layout Understanding via Human-in-the-Loop: A User Study on PDF-to-HTML Conversion for Long Documents

Chenyu Mao

(ABSTRACT)

Document layout understanding often utilizes object detection to locate and parse document elements, enabling systems that convert documents into searchable and editable formats to enhance accessibility and usability. Nevertheless, the recognition results often contain errors that require manual correction due to small training dataset size, limitations of models, and defects in training annotations. However, many of these problems can be addressed via human review to improve correctness. We first improved our system by combining the previous Electronic Thesis/Dissertation (ETD) parsing tool and AI-aided annotation tool, providing instant and accurate file output. Then we used our new pipeline to investigate the effectiveness and efficiency of manual correction strategies in improving object detection accuracy through user studies, including 8 participants, comprising a balanced number of four STEM and four non-STEM researchers, all with some background in ETDs. Each participant was assigned correction tasks on a set of ETDs from both STEM and non-STEM disciplines to ensure comprehensive evaluation across different document types. We collected quantitative metrics, such as completion times, accuracy rates, number of wrong labels, and feedback through our post-survey, to assess the usability and performance of the manual correction process and to examine their relationship with users' academic backgrounds. Results demonstrate that manual adjustment significantly enhanced the accuracy of document element identification and classification, with experienced participants achieving superior correction precision. Furthermore, usability feedback revealed a strong correlation between

user satisfaction and system design, providing valuable insights for future system enhancement and development.

Enhancing Layout Understanding via Human-in-the-Loop: A User Study on PDF-to-HTML Conversion for Long Documents

Chenyu Mao

(GENERAL AUDIENCE ABSTRACT)

With the development of technology, there is an increasing demand to make printed and scanned documents more accessible. Organizations such as universities and libraries have millions of valuable documents, including theses, dissertations, and research papers, which exist only in PDF, often as a scanned format. While these works contain valuable knowledge, they can be challenging to search through or access, especially for those with low vision. To solve this problem, we need computer systems that automatically recognize and convert different parts of these documents — like titles, headings, paragraphs, and figures — into more usable forms.

Our research focuses on improving how these document recognition systems work by combining computer automation with human expertise. While computers can process documents quickly, they sometimes need more training data for complex document layouts. We developed a web-based tool allowing people to review the computer's work and correct errors, such as mislabeled sections or missed elements. We conducted a detailed study with 8 participants who used our correction tool, to understand how effective this human-computer collaboration could be. We carefully measured several aspects of their experience: how many pages they annotated in a fixed amount of time, how accurate their corrections were, and how they felt about using the tool. We also used a post-survey to gather feedback about their experience with the tool.

The results were very encouraging. When humans reviewed and corrected the computer's

work, the accuracy of document recognition improved significantly. We found that participants could effectively identify and fix errors in the computer's output, especially when the tool was easy to use. Higher user satisfaction was strongly linked to how intuitive and straightforward participants found the correction process.

One useful finding was that this process creates a positive feedback loop. Every correction a person makes helps expand the training data available to the computer system, which means the system can learn from these corrections and gradually become better at recognizing similar elements in future documents, reducing the number of errors that need to be corrected over time. Our research offers insights into building advanced object detection systems incorporating computational efficiency with human review. The results boost the formulation of optimal strategies for developing user-centric interfaces and effective document repair operations. This work has practical implications for making academic and research documents more accessible to everyone, including those relying on screen readers or other assistive technologies. This research represents a step forward in making the vast knowledge of digital documents more accessible, searchable, and usable for all readers. By showing how humans and computers can work together effectively, we are helping to build better systems for preserving and sharing knowledge in the digital age.

Dedication

Dedicated to my beloved parents, Jianyang Mao and Jiqing Zhao, and to my dear sister Chenxin Mao

Acknowledgments

I would like to express my heartfelt gratitude to my advisor, Dr. Fox, for his constant encouragement and invaluable guidance throughout my study. Without his consistent and illuminating instructions, I would not have completed my thesis successfully. I also want to extend my deepest appreciation to my other mentor, Dr. Aman Ahuja. He introduced me to the fascinating world of research, and his influence has been profound. His expertise has not only been crucial in shaping my academic work but has also been a source of inspiration. I'm grateful to the Institute of Museum and Library Services (IMLS) for their funding support, through LG-256638-OLS-24 and LG-256694-OLS-24. It has been essential for our research, and I sincerely appreciate their contribution.

I would also like to express my sincere gratitude to my labmates, including Satvik Chekuri, Dr. Bipasha Banerjee, Sareh Ahmadi, and Pradyumna Upendra Dasu. It has been a privilege working alongside you during my studies. Special thanks go to my friends, Guanchen Wu, Aoqi Zeng, Yifei Wang, Junfei Wang, Xiao Guo, and Xuanang Zhao. They have accompanied me through the six years I spent in the United States. Whenever I was in need, they were always quick to offer a helping hand. Their friendship has been a constant source of support and comfort, making my time here more memorable and meaningful. I am truly fortunate to have them by my side, and I will always cherish the memories we've shared together.

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

ETDs Electronic Theses and Dissertations

FE Flawed Element

OD Object Detection

WL Wrong Label

Chapter 1

Introduction

Electronic Theses and Dissertations (ETDs) provide an essential store of specialized academic knowledge, offering significant advantages to many users within the scholarly community. Millions of ETDs are accessible online. Researchers frequently neglect their significant content due to three main obstacles: document length, the complexity of specialized information, and the absence of consistent formats throughout institutions and academic fields. This accessibility obstacle is especially critical for those dependent on assistive technologies like screen readers. The structural diversity of ETDs increases the difficulty in analyzing ETDs. Diverse academic disciplines lead to a variety of document components, causing conventional rule-based parsing techniques to be inadequate. Thus, mathematical texts feature equations, while computer science texts contain algorithms. The difference in formatting standards among institutions further complicates the design of universal parsing systems. These problems stress the necessity for more advanced, machine learning-driven methodologies capable of efficiently generalizing across various formats and domains. The latest developments in computer vision, particularly in object detection techniques, have shown significant potential for document parsing tasks [19]. Researchers have effectively used object detection algorithms to extract information from research papers; nevertheless, they have not substantially explored their application in larger, more complicated texts such as ETDs. The distinctive features of ETDs—length, specialized content, and structural complexity—require tailored systems capable of reliably identifying and extracting diverse

document pieces while preserving their contextual linkages.

This thesis tackles these problems by presenting a novel object recognition system specifically engineered to analyze extensive PDF documents such as ETDs and convert them into XML and then interactive HTML files. Our methodology transcends simple automated parsing by integrating a human-in-the-loop correction system. Typically, the quantity and quality of the training dataset determine the performance of the object detection system, which leads to the system consistently producing flawed elements. Besides, researchers have to put extra time into continuously labeling the elements in the images to ensure the performance of extracting texts from documents. However, our object detection system [2] provides a novel methodology that allows users to correct the flawed elements generated by the object detection model, and generate a flawless version based on user correction in real-time. This creative feature helps users to manually adjust the system's output, with these modifications incorporated into the training corpus to develop an ever enhancing model. Additionally, the system provides instant parsing requirements through this iterative improvement approach. Thus, the system continuously improves, to produce progressively more accurate XML files over time.

The purpose of this study is to explore and compare how individuals with diverse backgrounds interact with a human-in-the-loop artificial intelligence (AI) system designed for detecting, parsing, correcting, and extracting content from documents. By comparing accuracy, number of errors, and their relationship to document types and individual demographics, this study aims to: (a) investigate how individuals correct and validate system output relative to peer users and ground truth, and (b) analyze how individual demographic characteristics and document types affect user experience and interaction patterns. The results are expected to help enhance the users' understanding of the current model and its design interface, and to determine how the two should be optimally integrated to retrieve critical

information in document recognition systems. Ultimately, this study can help suggest how to make document recognition systems more effective and robust for different categories of users and to help advance the knowledge of how other users vary in their working with human-in-the-loop AI systems. After developing the system, we ran a pilot study with 2 participants and a user study with 8 participants recruited across the campus. The user study setting remained the same as our pilot study. We divided the 8 participants into four sets. The participants in a set worked on the same documents. Each participant had up to 3 hours to correct the labels and bounding boxes for their documents.

To assess the effectiveness of our methodology, we employed a mixed-methods study design that combines quantitative analysis of XML file accuracy with qualitative research on user interactions and experiences. Based on this approach, we proposed three hypotheses:

1. Manual correction will lead to the identification of many problems with the current system and its object recognition model.
2. Higher user satisfaction scores will be positively correlated with perceived ease of use.
3. Users with more significant academic experience and domain-specific knowledge will perform better at our task.

This thorough evaluation approach will assist in validating the technical efficacy of our system and its practical utility in enhancing the accessibility of ETD information to a broader audience.

Chapter 2

Review of Literature

2.1 Dataset

Our dataset has more than 500K ETDs covering the period from 1845 to 2020, but most are published after 1945 [2]. The labeling for the ETD objects was done using Roboflow [8], a free online labeling tool. Based on our analysis of a representative sample of 200 ETDs, we identified and labeled 24 categories that are commonly found in ETDs, such as paragraphs, figures, and tables. Table 2.1 presents these categories along with their frequency of occurrence in our labeled sample, which will serve as the target classes for our detection model [2].

Before starting the user study, we randomly selected a batch of 16 ETDs from a pool representing diverse disciplines, publication years, and universities, specifically for subsequent human-subject research as shown in Table 2.2. These documents were carefully chosen to ensure they contained a representative sample of errors that participants could work on during the study. We manually calculated the ground truth values for these documents to establish a baseline for evaluating participant performance in the correction tasks. The first four documents (Doc1-Doc4) were used in our pilot study, while all 16 documents were included in the main study.

Table 2.1: ETD Categories

Category	**#Instances [2]**
Title	439
Author	404
Date	338
University	309
Committee	282
Degree	279
Abstract Heading	169
Abstract Text	183
List of Contents Heading	512
List of Contents Text	1059
Chapter Title	2211
Section	9337
Paragraph	30359
Figure	6359
Figure Caption	5722
Table	2654
Table Caption	2213
Equation	5092
Equation Caption	3051
Algorithm	96
Footnote	5722
Page Number	24543
Reference Heading	313
Reference Text	2088
Total Objects	99859
Total Images	25073

2.2 Object Detection

Object detection remains an essential challenge in computer vision, with real-time detection important for different applications. The YOLO (You Only Look Once) family has developed significantly since its introduction, progressing from the original single-stage detection paradigm to increasingly sophisticated architectures [14]. Due to its architectural

Table 2.2: Document characteristics of the ETD dataset showing the distribution of STEM and non-STEM documents, along with their respective numbers of: pages, flawed elements (FEs), and wrong labels (WLs)

Doc	Type	Pages	FE	WL
1	Non-STEM	73	523	3
2	Non-STEM	50	38	4
3	STEM	34	15	4
4	STEM	52	65	5
5	STEM	39	86	4
6	Non-STEM	76	321	6
7	Non-STEM	55	602	2
8	STEM	55	205	14
9	Non-STEM	59	317	2
10	STEM	67	35	10
11	STEM	61	28	8
12	Non-STEM	80	29	41
13	STEM	74	0	5
14	STEM	42	10	2
15	Non-STEM	48	1	10
16	Non-STEM	74	36	10

advancements, YOLOv7 represents a significant progression in real-time object identification [17]. When trained from inception without pre-learned weights, the model attains an average precision of 56.8% on the MS COCO dataset, surpassing transformer-based and convolution-based object detectors across several circumstances [17]. YOLOv7 uses a convolutional neural network to simultaneously predict the entire image's bounding boxes and class probabilities, unlike previous object detection models that localize objects by analyzing certain portions of the image with elevated probabilities. The architecture consists of image frames processed through a backbone (a deep neural network with multiple convolutional layers for feature extraction), which is subsequently integrated into the neck (intermediate layers between the backbone and the head, serving as detectors to aggregate feature maps from various stages), ultimately identifying the bounding boxes. The backbones used in YOLOv7 are VoVNET, CSPVoVNET, ELAN, E-LAN. The neck in YOLOv7 uses FPN,

RFB, and PAN [17]. The detection happens in the head (dense prediction). Once the detectors predict localization and classification simultaneously, this layer is present at only one stage after detectors like YOLO, SSD, and RPN. Eventually, sparse prediction is for two-stage detectors, FRCNN and RFCN, which do the class probabilities for the model [7].

2.2.1 Document Layout Understanding

Document layout understanding and object detection are inherently connected in document analysis since layout understanding strongly relies on object detection algorithms to locate and identify elements in the document [10]. Object detection is the foundation of layout understanding by providing functionality, including predicting and identifying elements and recognizing hierarchical structures between elements. Such a relationship is demonstrated using a Transformer-based model, that achieves a remarkable element accuracy, 97.3% [16]. The result shows great potential for using object detection in document layout analysis.

2.2.2 AI-aided Annotation Tool

The AI-aided annotation system offers an innovative method for lengthy document analysis that uses pre-trained object detection models to create weak labels for documents. With the help of human verification, this tool could reduce annotation time by 2-3 times compared to the traditional method. The difference becomes significant when it comes to creating high-quality training datasets with minimum human effort [2]. The experiment's result proves that the framework increases the accuracy of extracting low-frequency elements from the documents, by up to a maximum of 20%, while maintaining resource efficiency [2].

2.2.3 ETD Parsing Tool

The ETD parsing tool framework proposed an end-to-end solution specifically for ETD parsing. The system is capable of extracting elements from ETDs and organizing them based on a predefined XML schema, creating a structured document representation [1]. The final experiment results show the framework's mean accuracy reached 85.3%, which provides a reliable technical foundation for ETD automation and digital applications [1].

2.3 Tools

2.3.1 Flask

The Flask framework is a lightweight web framework built on WSGI and written in Python, facilitating convenient and straightforward integration with our backend applications [11]. WSGI, or Web Server Gateway Interface, delineates the communication protocol between the web server and the web application [15]. Its micro-framework architecture renders it adaptable and highly scalable, offering various third-party libraries. We can integrate it with the necessary Python libraries during project development. Furthermore, implementing additional web features and functionalities would be straightforward. Additionally, we require front-end technologies, specifically HTML, CSS, and JavaScript. HTML will provide the fundamental framework of the website, CSS will govern its aesthetic presentation, and JavaScript will enhance user engagement.

2.3.2 Pdf2image

Pdf2image is a valuable Python module that can transform Portable Document Format (PDF) files into PIL (Python Imaging Library) image objects [3]. The function yields a PIL image object corresponding to each page of the specified PDF file located in the provided local directory. This function is executed as a preprocessing step to obtain the set of page images that will serve as input for the object detection model. Additionally, we will store the page pictures for ETDs in the repository, which can be used for further model training.

2.3.3 React-Image-Annotate

React-Image-Annotate is a library that primarily focuses on building annotations and labels in pictures [18]. It is usually used for machine learning and computer vision applications because the model training process mainly relies on the quality of these annotations. To address this problem, react-image-annotate provides diverse annotation methods, such as bounding boxes, polygons, and point annotations. These diverse annotation methods make the whole computer vision annotation work versatile. In addition, it provides customized keyframe integration and configurable hotkeys, which improves user experience and flexibility in operations. Due to different requirements in popular machine learning frameworks, the library could output standard annotation file formats, such as COCO, VOC, and YOLO, and these formats can be directly used in model training without an additional format transformation [18]. The library's utilization in academic research and industrial applications illustrates its sophistication and dependability in visual data annotation.

2.4 Human-in-the-loop

Human-in-the-loop plays a significant role in machine learning pipelines, such as data extraction, integration, cleaning, annotation, iterative labeling, model training, and model inference. One noticeable challenge for a human-in-the-loop system is reducing the cost while maintaining the quality of the model. With a limited budget, researchers tend to label the data iteratively, and improve the model performance by adopting techniques like active learning [5, 6]. Active learning can enrich human labeling in two scenarios: front-up and iterative. In the front-up scenario, users focus on latency instead of cost, so they are prone to select some representative labels to train an initial model.

In contrast, in the iterative scenario, users select another batch of labels to train the model iteratively, until the budget is used up. During the active learning process, there are two strategies for choosing the training dataset: Uncertainty and MinExpError [5, 6]. Uncertainty represents the samples that the current model is most uncertain about, while MinExpError represents the samples that can maximize the improvement of the model's future performance.

In our research, we aim to build a model that can accurately identify and parse all elements in ETDs for HTML rendering. We adopt an iterative active learning approach to optimize our model's performance. Initially, we recruited undergraduate students to manually label a small set of ETDs, which served as the foundation for training our initial object detection model. Following the MinExpError strategy discussed above, we iteratively selected a batch of ETDs where our system produced errors. These challenging cases were then manually labeled by the participants and used to further enhance the performance of our current object detection model.

Chapter 3

Overview of Related Systems

3.1 AI-aided Annotation Tool

The AI-aided annotation system inputs a document and outputs multiple text files that save the coordinate information for contained elements. The process can be concluded in three steps: data sampling, weak label generation, and manual verification. In the first step, as shown in Figures 3.1 and 3.2, the framework requires the user to submit the documents in a single PDF of a single page. Then, it will further require uploading the categories and the model, as presented in Figures 3.3 and 3.4. After that, as is illustrated in Figure 3.5, the system generates the label and bounding box for each category in the document, and users need to verify labels and bounding boxes to prevent inaccurate judgments made by the system. Finally, as illustrated in Figure 3.6, users have the flexibility to specify both the train-validation split ratio and the preferred format for coordinate data export.

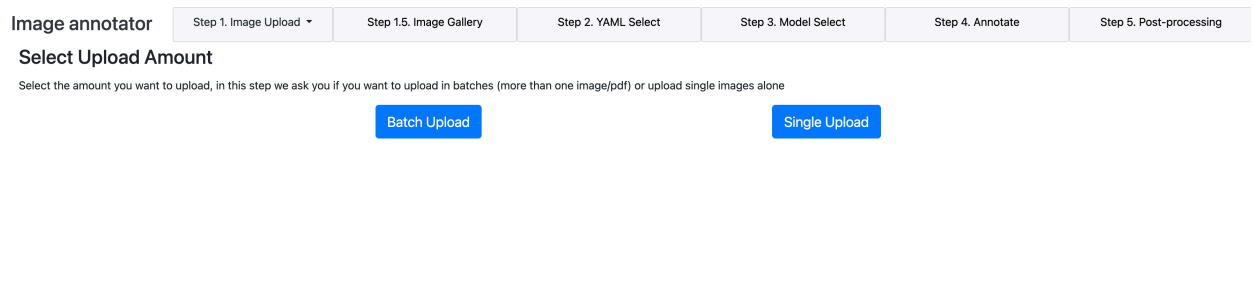


Figure 3.1: AI-aided Annotation Tool: Home page for selecting upload mode

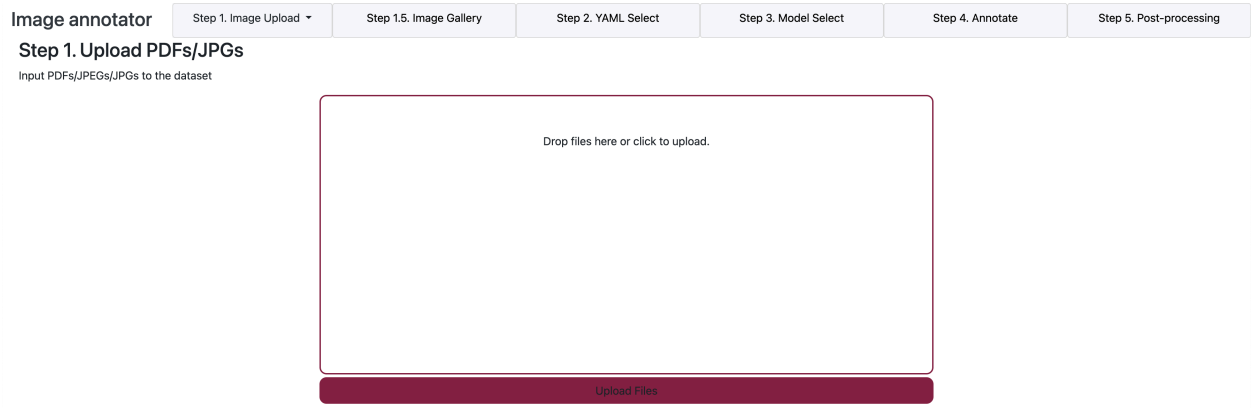


Figure 3.2: AI-aided Annotation Tool: Document upload page for submitting PDFs and image files

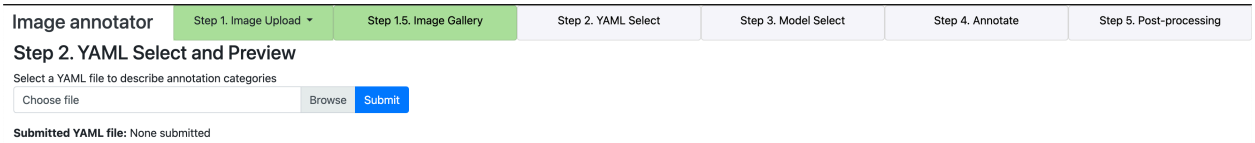


Figure 3.3: AI-aided Annotation Tool: Document upload page for uploading categories

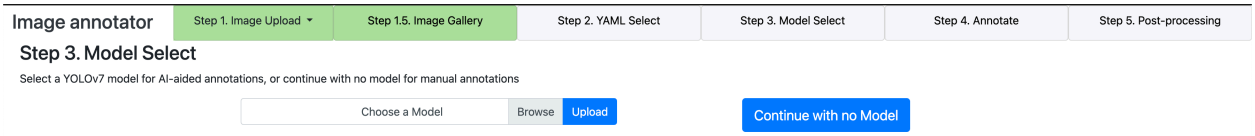


Figure 3.4: AI-aided Annotation Tool: Document upload page for uploading object detection model

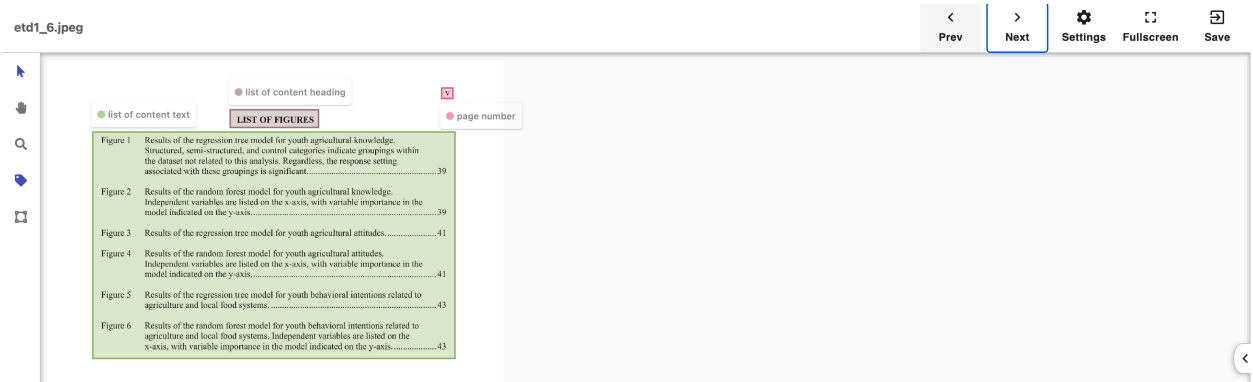


Figure 3.5: AI-aided Annotation Tool: Document annotation page for correcting wrong labels and inaccurate bounding boxes

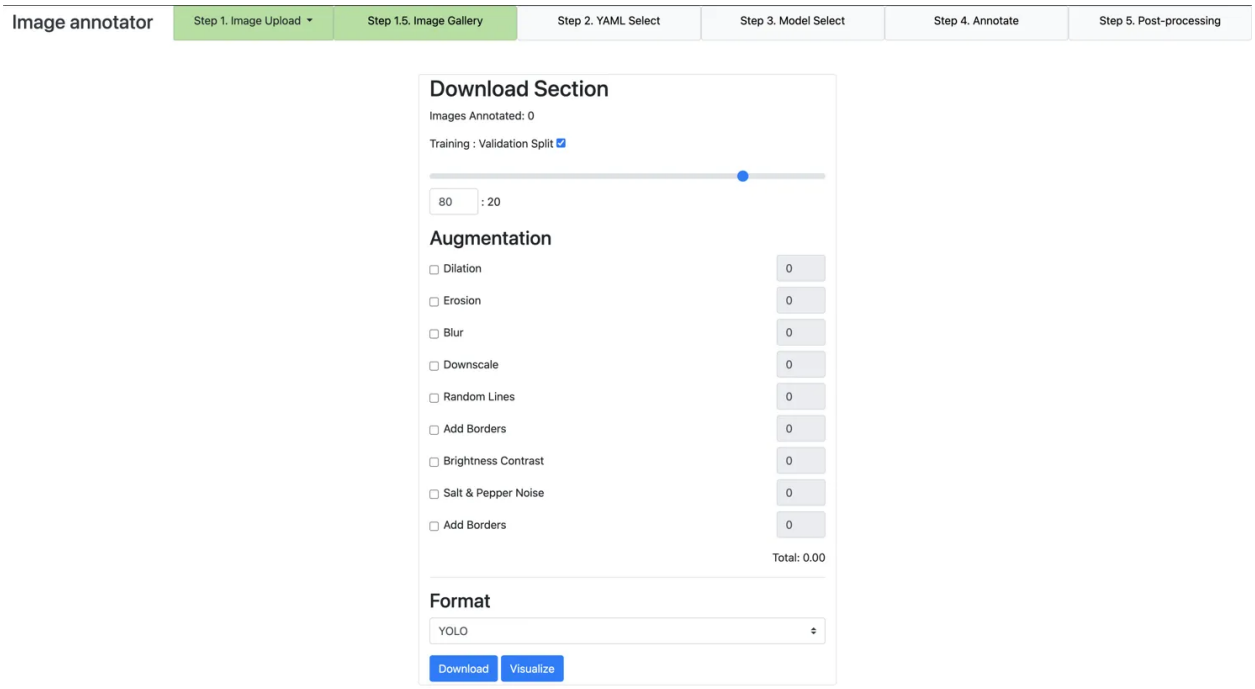


Figure 3.6: AI-aided Annotation Tool: Post-processing configuration page displaying data split ratio, augmentation options, and export format settings

3.2 ETD Parsing Tool

The ETD parsing tool processes documents by extracting and organizing their textual and visual content into a hierarchically-structured XML output while preserving the semantic relationships between elements. The system's home page features a dual-model selection interface for object detection, as illustrated in Figure 3.7. Users can then proceed to upload a document through either browsing their local device or utilizing the drag-and-drop functionality within the designated dashed rectangular area, as depicted in Figure 3.8. In the preprocessing stage, the framework splits PDF documents into page images. After that, a pre-trained object detection model will identify key elements on each page and categorize them into text-based (paragraphs and metadata) and image-based (figures, tables, and equations) elements. Ultimately, the system extracts textual content from the text-based elements via the OCR tool and establishes an association with the image-based elements.

Figure 3.7: ETD Viewer: A tool for choosing between object detection models, featuring options for Detectron2 and YOLOv7

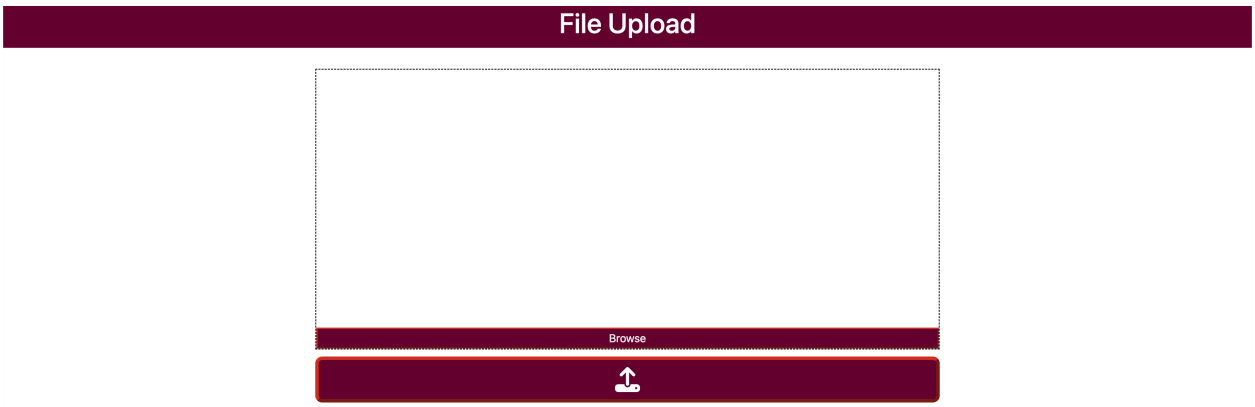


Figure 3.8: ETD Viewer: File upload interface with a dotted border drop zone and a 'Browse' button, designed for users to upload files either by dragging and dropping or selecting through the file browser

The processed results are then exported to an XML file and rendered as an HTML page, as illustrated in Figure 3.9.

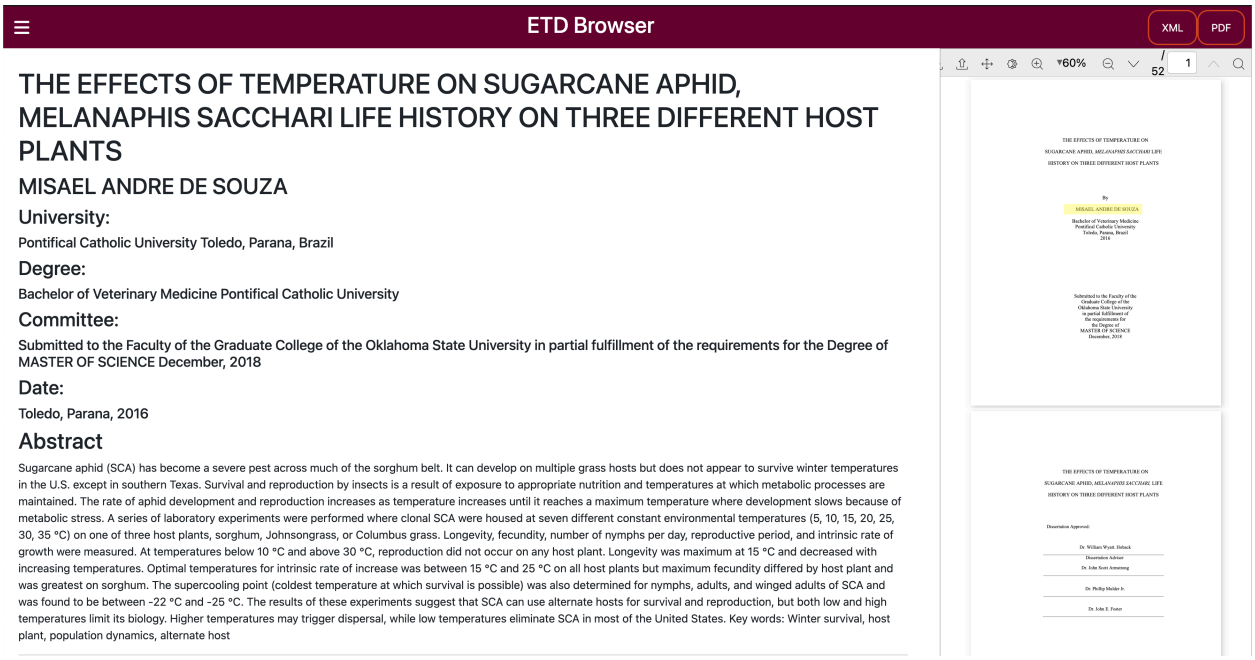


Figure 3.9: ETD Viewer: A synchronized split-view document viewer that displays editable text content on the left and the original PDF on the right, with interactive highlighting that shows corresponding PDF sections in yellow when text is selected

Chapter 4

Human-in-the-loop Document Parser

4.1 Background

Although the ETD parsing tool provides an instantaneous framework capable of parsing the document and providing an organized and interactive representation, some features still need to be extended. The primary challenge lies in the high dependency on high-quality object detection models. The performance of the system deteriorates significantly when it comes to parsing documents with a layout that does not exist in the training dataset. This defect becomes obvious when processing documents with diverse formats, disciplines, and styles of figures and tables. Under the circumstance of a limited quality training dataset, the parsing tool might give a confusing and disorganized output, which then is rendered in an HTML page. Users inevitably need to manually check the elements from the original document again, thus making our system ineffective and inefficient. Such inconsistent accuracy performance across different ETDs leads to an increasing demand for human intervention. With the help of the AI-aided annotation tool, inaccurate bounding boxes and wrong labels are fixed by human checks, and the results of manual adjustments for every page are saved in text files for future model training. For each page, each element generated by the system will have a unique bounding box coordinate, and the representation of this coordinate is in a format called YOLO coordinate format [2]. It is foreseeable that these corrections can serve a dual purpose; they can not only be used for model training but also to generate accurate

```
19 0.5374457056382123 0.24222104852849788 0.7561234956629136 0.34002661965110087
19 0.5447277203728171 0.7036244340376421 0.7776914170209099 0.44204001686789773
2 0.4992990471335018 0.45509883533824574 0.15348902085248162 0.062315757057883524
18 0.8753709142348346 0.04540259794755415 0.028914651309742648 0.01930990392511541
2 0.4984048102883732 0.4714897294477983 0.1181353759765625 0.027596380060369317
2 0.4958096313476562 0.44068226207386363 0.1279016472311581 0.027795521129261362
```

Figure 4.1: Example of an Annotation File. This file consists of five columns: The first column represents the category index, while the remaining four columns contain the YOLO bounding box coordinates.

bounding box coordinates for the ETD parsing tool to extract information.

The frontend file browsing functionality is implemented using the react-image-annotate library. In this library, bounding boxes are represented by their four corners: top-left is $(x_{\min}, y_{\min})$, top-right is $(x_{\min}, y_{\max})$, bottom-left is $(x_{\max}, y_{\min})$, and bottom-right is $(x_{\max}, y_{\max})$. However, the YOLO model training process requires the YOLO coordinate format, which uses four different values to represent a bounding box; they are x_{center}, y_{center}, $width_{\text{norm}}$ and $height_{\text{norm}}$. The equations for coordinates of YOLO format are as follows:

$$x_{\text{center}} = \frac{x_{\min} + x_{\max}}{2 * \text{image\_width}}, \quad y_{\text{center}} = \frac{y_{\min} + y_{\max}}{2 * \text{image\_height}}$$

$$width_{\text{norm}} = \frac{x_{\max} - x_{\min}}{\text{image\_width}}, \quad height_{\text{norm}} = \frac{y_{\max} - y_{\min}}{\text{image\_height}}$$

To convert YOLO format to traditional coordinates, we can derive the equations through the following steps:

For the x-coordinates:

$$x_{\text{center}} = \frac{x_{\min} + x_{\max}}{2 * \text{image\_width}}$$

$$2 * x_{\text{center}} * \text{image\_width} = x_{\min} + x_{\max} \tag{1}$$

$$width_{\text{norm}} = \frac{x_{\max} - x_{\min}}{\text{image\_width}}$$

$$width_{\text{norm}} * \text{image\_width} = x_{\max} - x_{\min} \tag{2}$$

From equation (2):

$$x_{\max} = x_{\min} + width_{\text{norm}} * \text{image\_width} \tag{3}$$

Substituting (3) into (1):

$$2 * x_{\text{center}} * \text{image\_width} = x_{\min} + (x_{\min} + width_{\text{norm}} * \text{image\_width})$$

$$2 * x_{\text{center}} * \text{image\_width} = 2x_{\min} + width_{\text{norm}} * \text{image\_width}$$

$$2x_{\min} = 2 * x_{\text{center}} * \text{image\_width} - width_{\text{norm}} * \text{image\_width}$$

$$x_{\min} = (x_{\text{center}} - \frac{width_{\text{norm}}}{2}) * \text{image\_width}$$

Similarly, substituting the derived $x_{\min}$ back into (3):

$$x_{\max} = (x_{\text{center}} + \frac{width_{\text{norm}}}{2}) * \text{image\_width}$$

Following the same process for y-coordinates:

$$y_{\min} = (y_{\text{center}} - \frac{height_{\text{norm}}}{2}) * \text{image\_height}$$

$$y_{\max} = (y_{\text{center}} + \frac{height_{\text{norm}}}{2}) * \text{image\_height}$$

Therefore, the complete conversion from YOLO format to traditional bounding box coordi-

nates is:

$$
\begin{aligned}
x_{\min} &= (x_{\text{center}} - \frac{width_{\text{norm}}}{2}) * \text{image\_width} \\
x_{\max} &= (x_{\text{center}} + \frac{width_{\text{norm}}}{2}) * \text{image\_width} \\
y_{\min} &= (y_{\text{center}} - \frac{height_{\text{norm}}}{2}) * \text{image\_height} \\
y_{\max} &= (y_{\text{center}} + \frac{height_{\text{norm}}}{2}) * \text{image\_height}
\end{aligned}
$$

4.2 Architectural Overview

We now introduce the proposed framework for parsing long documents into interactive HTML pages. The architecture of this framework is illustrated in Figure 4.2. The different modules can be roughly divided into six steps, shown in the figure, and explained in the following six sub-sections.

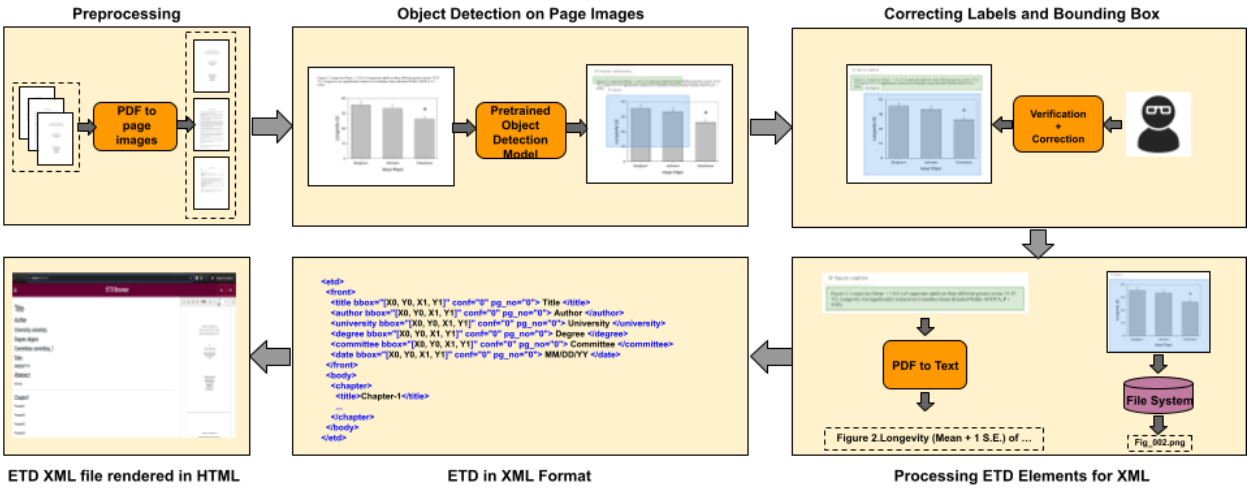


Figure 4.2: Architecture of the proposed PDF to XML parsing framework

4.2.1 Data Preprocessing

Our system is designed mainly for parsing extensive scholarly publications, utilizing the PDF form of the document as input. The input file is transformed into separate page images (.jpg format) via Python-based PDF tools like pdf2image [3]. The page images are subsequently inputted into the Element Extraction module for additional processing. The first page of the framework provides a straightforward interface with two buttons for file uploading, as shown in Figure 4.3. In this step, users must choose the amount they want to upload; if the file contains multiple images, they need to use the Batch Load button; otherwise, they use the Single Upload button.

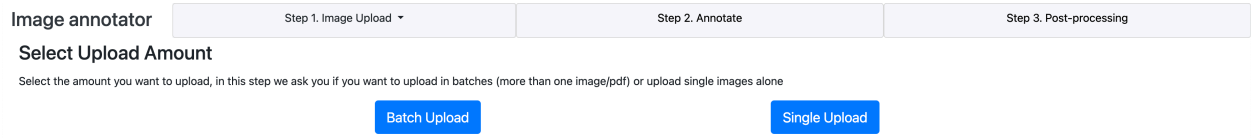


Figure 4.3: File uploading page used to upload multiple images or a single image

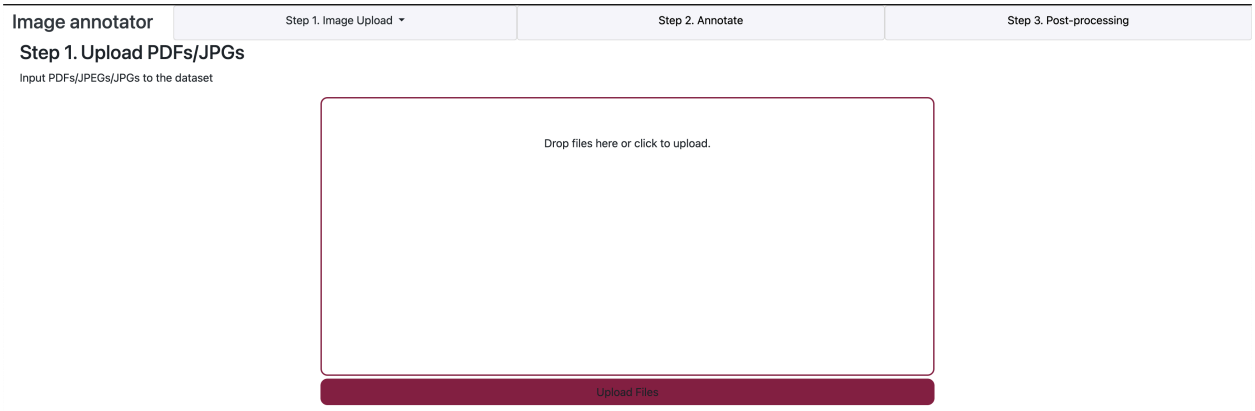


Figure 4.4: File uploading page used to browse a file and upload it

4.2.2 Element Extraction via Object Detection

The object detection module serves as the backbone of our framework. It takes an individual image as an input and uses an object detection model such as YOLO7 to extract the elements. The output of this inference procedure is a list of elements. Each element contains some descriptive data—such as bounding box coordinates in the YOLO format, category index, and page index—which will be used later for human verification and training dataset enrichment. The framework will iterate over all pages in the uploaded document, each yielding a corresponding list of elements. There are no changes to the model parameters during this inference process, which employs object detection to identify and describe each of the objects on each page.

4.2.3 Human Verification and Correction

Before the extracted elements are sent to assemble the final XML file, they must pass the human verification and correction round. The bounding boxes and labels inferred by the model are not always precise. As shown in Figure 4.5, two overlapped bounding boxes cover the figure caption, representing two objects. Besides, the system creates a bounding box that does not fully cover the figure. Thus, humans can intervene by using the toolbar provided on the left of the page.

That toolbar has five icons. The first icon, an arrow, enables users to select a single bounding box. Then, a selection box will pop up, allowing the user to browse through the element category options, and replace the system assigned category with the correct one. Further, when the mouse hovers over the four corners of the bounding box, it transforms into a drag handle, allowing the user to adjust the size and corners of the box. The second icon, which looks like a hand, enables users to move the position of the document and makes it easier

to make locational corrections. In addition, the third (magnifying glass) icon allows users to zoom in and out of the document, more precisely adjusting the bounding box coverage. The fourth icon serves as the 'show-mask' tool, which enables users to toggle the visibility of annotation labels, facilitating easier review and management of labels. Last but not least, the fifth (bounding box) icon can create a new bounding box, which can be used to capture elements missed by the model inference. At the top right, below "Step 3. Post-processing", is another toolbar, for page browsing. The Prev and Next buttons let the user move to the last and next page, while the Fullscreen button can turn the browser into full-screen mode. After users finish the verification and correction of every page in the document, they need to record the changes by clicking the Save button.

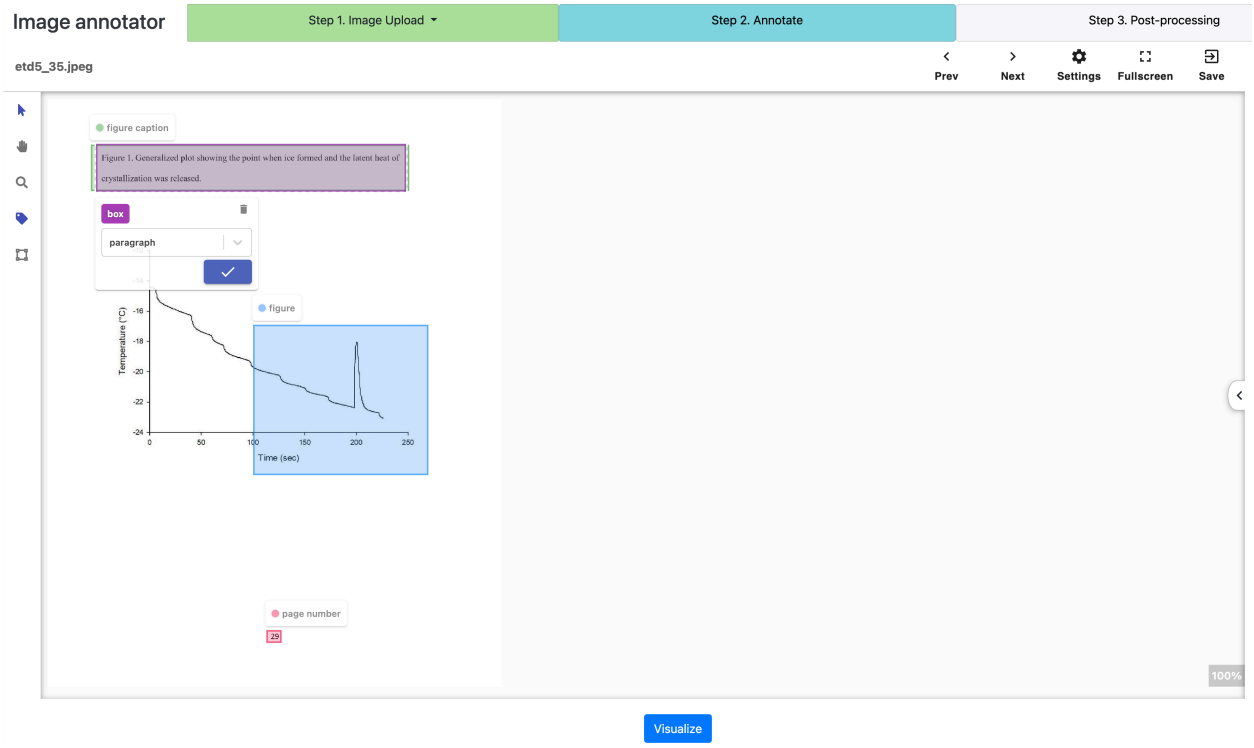


Figure 4.5: Annotation page used to correct bounding boxes and labels

4.2.4 Structuring Elements into XML

After extracting all the elements from the document, the system outputs the representation. We broadly classify the elements into two categories. The first category includes image-based objects, such as figures, tables, and equations. All of them will be cropped based on the coordinates, and stored in the file system. The second category is text-based objects—including metadata, paragraphs, etc.—from which text is extracted through further processing.We used some off-the-shelf Python libraries to extract plain text from text-based objects. Since newer PDF files are born digital, we use a Python library called pymupdf to extract text from elements in those documents. Since older PDF documents have images of pages, we used an optical character recognition (OCR) library named pytesseract to identify text [13] [9]. After collecting all the text and image elements from the document, we organized them according to our XML schema.

4.2.5 Visualization of XML

The output XML file currently has a suboptimal presentation. We want to explore ways to improve its accessibility, such as converting it into a more easily rendered format like HTML. HTML offers greater flexibility and improved usability. To achieve this, we have chosen to reconstruct the XML format into an HTML version using the Python library, ElementTree [1]. Users can visualize the XML output by clicking the Visualize button on the bottom of the annotation page (see Figure 4.5), which leads you to the interactive HTML page, as shown in Figure 4.6. This interface features a three-panel layout: a navigation sidebar, a document viewer, and a PDF viewer. The comprehensive navigation sidebar is particularly valuable for accessing lengthy Electronic Theses and Dissertations (ETDs), allowing users to locate and quickly jump to specific content of interest. Additionally, the hierarchical

navigation structure provides a detailed document breakdown, organizing content by both sections and subsections, enhancing the overall browsing experience. The document view in the middle offers the content of the document. Metadata content, including title, author, university, etc., is provided at the start of the document view. Below this information are paragraphs, figures, and tables extracted from the document. By clicking the document text, users can reference the corresponding location in the PDF, which is shown highlighted in a yellow bounding box. Additionally, two buttons in the top right corner allow users to download the parsed XML files and the original PDF document.

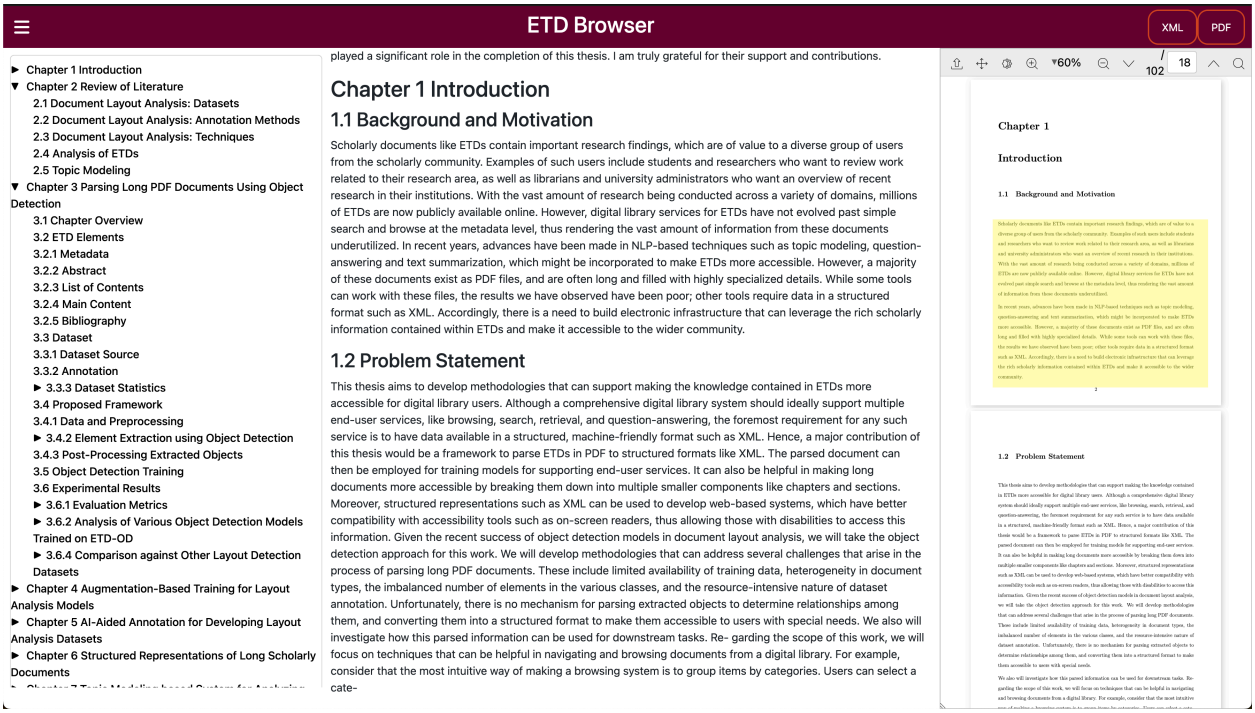


Figure 4.6: Visualization page of the output XML file

4.2.6 Post-processing Page

After users finish the annotation work, the framework will lead them to the post-processing page, as shown in Figure 4.7. This page lets users download the annotation files for future

training dataset enrichment, while the drag bar allows users to adjust the proportion of training and validation annotation files. In the future, if researchers are interested in another model that requires a different coordinate format, they can choose the file format by selecting the corresponding option. Previously, we mentioned that the framework also provides an updated version of XML output based on manual correction, and users can review the updated XML output rendered in the HTML page by clicking the PostVisualize button.

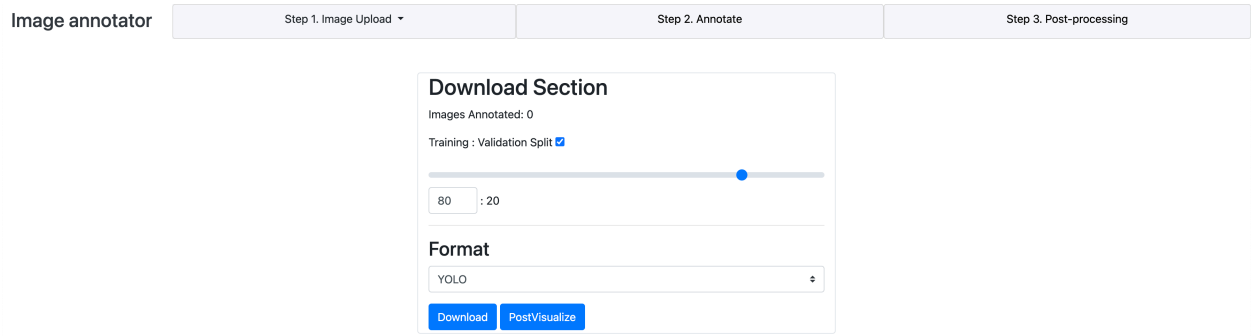


Figure 4.7: Post-Processing page used to download annotation data

Chapter 5

User Study Design

This chapter explains the design of the user studies, including the pilot and main study.

5.1 Hypotheses

We conducted both a pilot and a main user study to explore and compare how individuals interact with a human-in-the-loop AI system designed to detect, parse, and correct information from ETDs. Based on our experience and intuition, and to provide guidance for future related work, we formulated three interrelated hypotheses (given in the first chapter, repeated here for convenience):

1. Manual correction will lead to the identification of many problems with the current system and its object recognition model.

2. Higher user satisfaction scores will be positively correlated with perceived ease of use.

3. Users with more significant academic experience and domain-specific knowledge will perform better at our task.

5.2 Evaluation Criteria

Before we conducted the main user study, we submitted our study plan, description, informed consent, and recruitment materials to the Institutional Review Board for approval; the corresponding approval letter can be found in Appendix D. We planned to recruit a balanced number of individuals with STEM and non-STEM backgrounds.

Each individual needs to verify and correct the wrong labels and inaccurate bounding boxes in a set of ETDs that our system outputs. Each set of ETDs includes two STEM-related ETDs and two non-STEM ETDs to facilitate comparisons, which allows us to examine the relationship between the academic backgrounds of the participants and their effectiveness and efficiency in working within and outside their fields of expertise.

During a participant session, the system will record all user actions, including updated bounding box coordinates for each page, the number of incorrect labels, and the amount of time spent completing the task. Meanwhile, we calculate the number of wrong labels and flawed elements generated by the system to compare them with the numbers after user corrections. This comparison allows us to evaluate the effectiveness and efficiency of the human-in-the-loop system.

Paired t-test is a method to determine whether there is a significant difference between two paired samples. However, since our data does not fit the normal distribution due to small sample size, we adopted the Wilcoxon signed rank test. We calculated the average improvement percentage for each group by taking the difference in the values from before and after human correction, and dividing this difference by the initial value before human correction. Then, we could see the average improvement percentage and test our first hypothesis.

A Pearson correlation matrix allows us to examine the correlation between variables. We collected user feedback through participant ratings across five dimensions: system learnability,

toolbar error functionality, task completion independence, error correction ease, and system stability. System learnability indicates whether users can master the system after tutorials, and toolbar error functionality assesses if system features effectively support correction work. Task completion independence measures users' ability to complete correction tasks without additional assistance from researchers, while error correction ease reflects the simplicity of carrying out correction operations. System stability indicates whether crashes occur during correction work. Subsequently, we calculated the Pearson correlation coefficients to characterize the relationships between user satisfaction and perceived ease of use, considering error correction ease and the other evaluated dimensions, and thus test our second hypothesis.

Prior to beginning the formal correction work, participants completed a demographic survey which included assessing their familiarity with ETDs on a five-point scale (extremely familiar, very familiar, moderately familiar, slightly familiar, and not at all familiar). We analyzed the average improvement percentages across different familiarity groups to investigate potential correlations between ETD familiarity and correction work efficiency. We also examined other demographic aspects, such as academic experience and domain-specific knowledge.

The number of wrong labels reflects the following scenarios: missing bounding boxes and wrong labels (i.e., mis-classification of an element). The number of flawed elements represents the total count of individual words affected by inaccurate bounding-box detection during the annotation step. These inaccurate bounding boxes lead to various XML tagging errors, which can be categorized into three types:

1. Missing words: When a bounding box fails to fully encompass a content region, the corresponding XML tags may be incomplete. For example, if the bounding box of a title only partially captures the title text, some words will be omitted in the XML file, resulting in incomplete content in the final HTML output.

2. Duplicated words: When overlapping bounding boxes are detected, the same content may be tagged multiple times in the XML file. This results in duplicate words appearing in the rendered HTML.

3. Irrelevant words: When a bounding box extends beyond its intended content area, it may capture unrelated content. For instance, if the bounding box of a title incorrectly includes the author section, the name of a author will be wrongly tagged as part of the title in the XML file.

Each word affected by these bounding box-induced tagging errors is counted as a flawed element. This includes both textual elements and figures where the bounding boxes fail to accurately define their boundaries. By comparing these numbers, we could analyze users' behavior and their understanding of different sections of ETDs. In addition, our XML template provides metadata of ETDs at the top, including title, author, degree, etc. These metadata fields are crucial for some document retrieval systems, as they serve as filtering criteria when users search for specific documents. Inaccurate metadata could compromise the effectiveness of the document filtering function. However, metadata retrieved by previous ETD parsing tools frequently contains unrelated descriptive phrases. For example, instead of just extracting "Master of Science," the output might include the verbose preamble: "A thesis submitted to the Graduate Faculty of Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of Master of Science.", as shown in Figure 5.1. In such cases, we classify the preamble text (e.g., "A thesis submitted to the Graduate Faculty of Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of") as flawed elements, and each word in the "extra" phrase is counted as a flawed element. Since ETDs, due to differences in schools and disciplines, typically feature very diverse styles in terms of metadata and figure presentation, our current object detection model is prone to make mistakes in several places, including metadata,

figure, figure caption, table, and table caption. Our system usually makes mistakes in identifying figures and figure captions, since, normally, figure captions are written in bold text and will have extra indent spaces above and below. Thus, if the caption of the figure is very closely placed relative to the figure and is not written in bold text, our system will not be able to distinguish them, as shown in Figure 5.2, where the figure numbered 3.1 is labeled as paragraph. Another common situation is that the figure and the figure caption are arranged horizontally instead of the common vertical way, as shown in Figure 5.3, where our system accidentally takes the paragraph under the figure as the figure caption and treats the figure and real figure caption as a single figure label. This circumstance will result in two wrong labels, one missing block of figure caption text, and one duplicated block of paragraph text. Among all the ETDs selected for the user study, metadata and figure errors were the most prevalent.

Before starting the user study, we randomly selected a batch of ETDs from a pool representing diverse disciplines, publication years, and universities. We then calculated their ground truth values and ensured there were some errors so that participants could work on them. Our system extracts text and images based on bounding box coordinates and generates XML files, which are subsequently rendered as HTML web pages. The ground truth was determined by conducting a detailed comparison between these rendered HTML pages and their original PDF documents. Any mislabeled bounding boxes or inaccurate coordinates directly affect the final HTML rendering, causing content misplacement in the output. For instance, if a regular paragraph is incorrectly labeled as an abstract, its text content would appear in the abstract section of the rendered HTML page instead of its original location in the document structure. As detailed in Section 2.1, we had previously established a collection of ETDs with their corresponding ground truth values. From this pre-validated dataset, we selected the first four documents for our pilot study, while the complete set of 16 documents

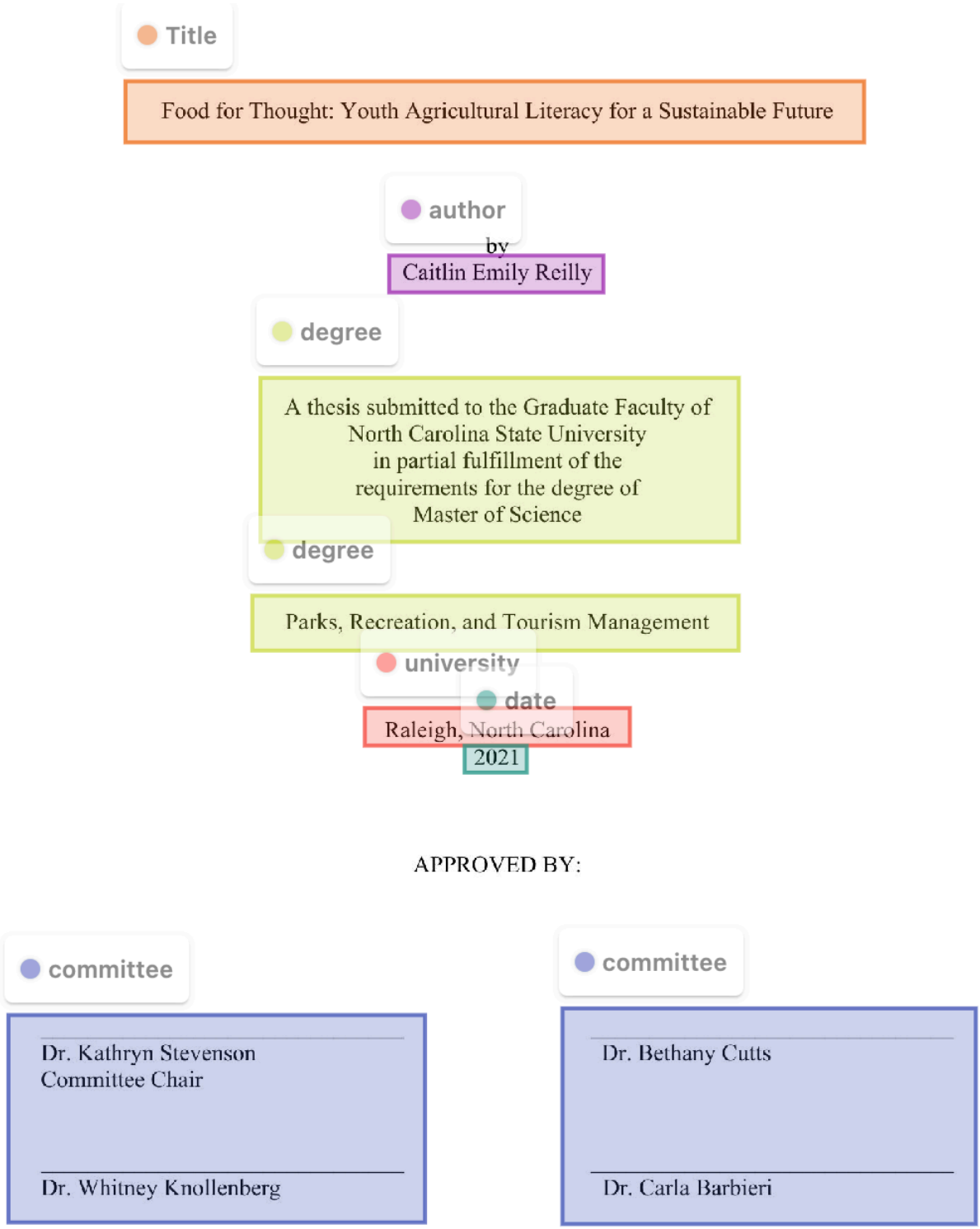


Figure 5.1: A common metadata error in the experiment ETD produced by the system

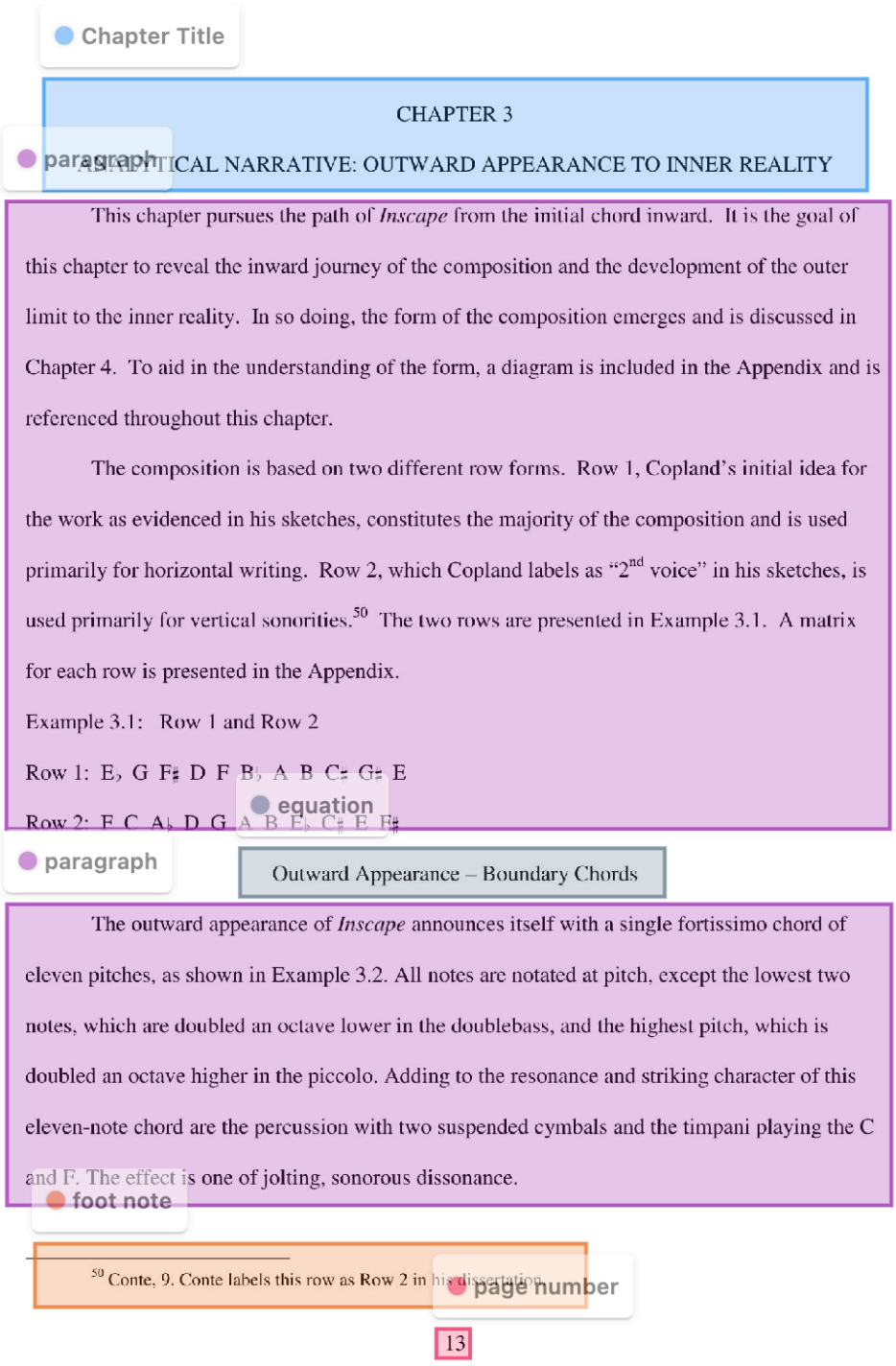


Figure 5.2: A common error caused by the lack of bold font in figure captions and the absence of extra indent space above and below the figure, making it difficult to distinguish between the figure and its caption.

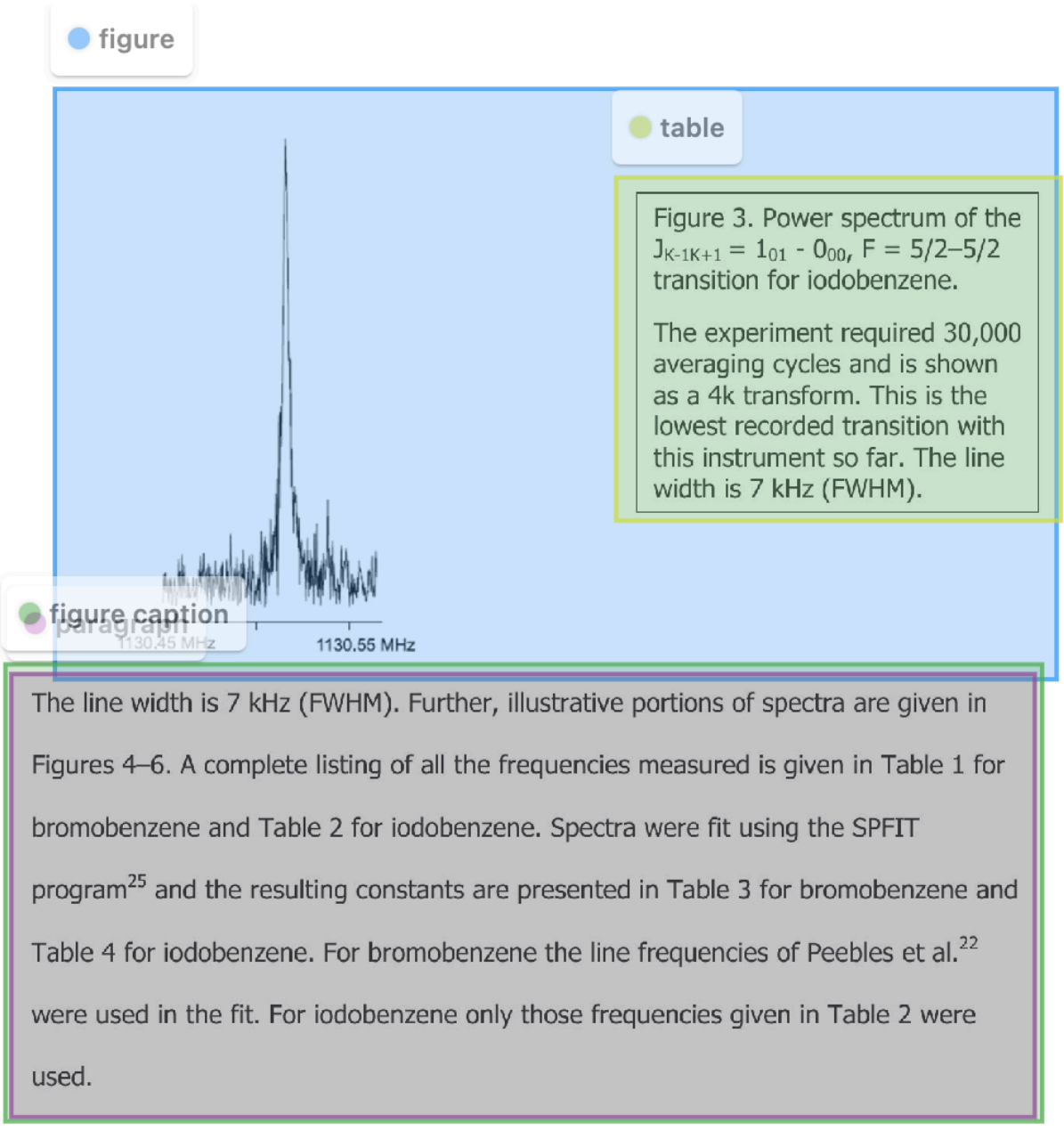


Figure 5.3: A common error caused by a novel format that makes it difficult to distinguish between figures and figure captions

was subsequently utilized in the formal main study.

5.3 Study Procedures

The study was conducted in a dedicated research laboratory room with a closed door, ensuring a quiet and controlled environment without external disturbances. Each participant worked individually on a Linux-based computer equipped with Internet connectivity and a 27-inch monitor. The experimental setup was designed to minimize distractions and maintain consistency across all sessions. The participant's task is to use the homegrown system to correct the wrong labels and inaccurate bounding boxes in an XML file, which will be displayed in a rendered interactive HTML page that our computer software generates after trying to identify the various document elements (e.g., paragraph, equation, page number). We will not collect user activity data other than when they are using our homegrown web-based system, which will be set up for them so they don't need to log in or provide any personal information aside from general demographics. During this process, there will not be any video or audio capture. However, our system will record all user actions related to our system as they complete the task; we ask them to stay focused on the given task. In addition, participants were instructed not to open any other program or application on the computer during the process, but were informed that they could take breaks at any time if needed. A detailed breakdown of our task is provided below.

1. Participants will be guided to the VT version of the QuestionPro platform to complete a survey where they should carefully read the background information about this research [12]. They can move forward only after they give consent, as shown in Appendix A.

2. After providing their consent through QuestionPro, participants will be asked to answer

some demographic questions, such as age, major, and category (undergraduate or graduate student, member of faculty or staff, researcher). If they meet the specified criteria, they will then receive detailed instructions on how to use the system.

3. Next, each participant will be assigned a set of 4 ETDs through a controlled random assignment process. To maintain our experimental design balance (one STEM and one non-STEM student in each group), the ETDs are randomly distributed while ensuring this group composition requirement is met. Participants will use the toolbar, mouse, and keyboard to correct any errors (wrong labels or inaccurate bounding boxes) in the XML files, and save the corrected XML result by clicking the "Save" button on the toolbar. Though focused on high-quality work, they should try to complete the checking of all 4 ETDs within the 3 hours allowed. It is also fine if they complete the checking in less than 3 hours. After that, they will provide feedback.

4. Finally, after participants finish the annotation of as much as they can in the given set of ETDs, participants will be asked to return to the QuestionPro form to provide feedback as well as ratings based on key usability evaluation criteria (such as ease of use, satisfaction, and learnability) using a Likert scale [12].

5.4 Participants

Since participants are tasked with correcting errors in a set of ETDs, individuals with extensive experience working with theses and dissertations are given high priority, so we target recruiting undergraduate students with research experience, graduate students, faculty members, and researchers. Faculty members across various departments and disciplines are included to ensure diverse perspectives and expertise. Since the participants must come to the lab and finish the user study, it requires them to be at Virginia Tech. Recruitment emails

will be sent to potential participants via university-approved mailing lists and direct emails. These emails will include a brief description of the study, eligibility criteria, participation instructions, and a statement emphasizing participation is voluntary and individuals should not feel pressured to join the study, even if the recruitment material is shared by someone we directly contacted. For follow-up communications, we will limit contact to a maximum of one initial email and no more than two follow-up emails.

Chapter 6

User Study

6.1 Pilot User Study

Before we proceeded to our full-scale user study, we conducted a pilot study to examine the feasibility of the study workflow and identify any potential issues. We recruited two participants from across the campus. Each participant was allocated three hours to complete the correction tasks using a subset of documents (specifically, the first four documents) from our pre-validated dataset, as described in Section 2.1.

6.2 Pilot User Study Results

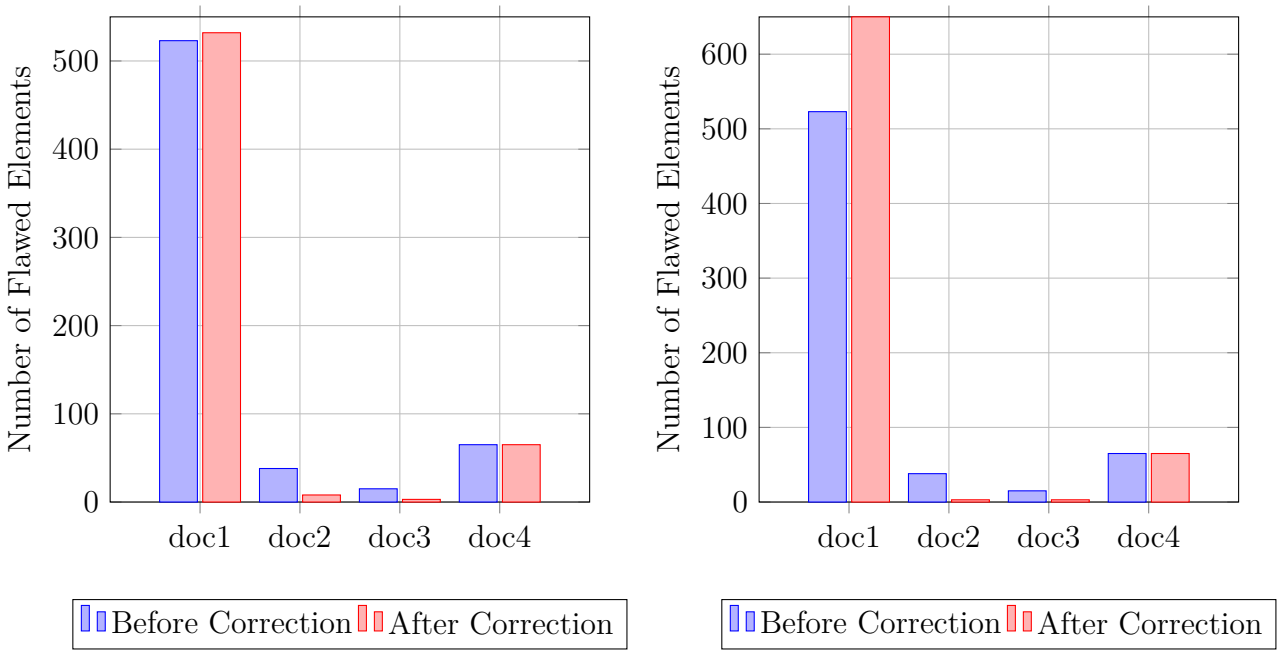
After counting the number of errors in the XML file rendered by the HTML page, we made a few tables and graphs that are given below, representing the output properties before and after human correction. First, we count the number of errors generated by our system and also include the number of pages and type of documents, as shown in Table 6.1. After that, we plot two bar charts to give a more straightforward presentation of the result, so we can clearly see how many changes they made in the experiment.

We examine the effectiveness of our system in the number of flawed elements. As presented in Figures 6.1a and 6.1b, the number of flawed elements in documents 2 and 3 decreased sig-

Table 6.1: Document Properties and Errors

Doc	Type	Pages	FE	WL
1	Non-STEM	73	523	3
2	Non-STEM	50	38	4
3	STEM	34	15	4
4	STEM	52	65	5

nificantly, indicating participants made correct adjustments to the bounding box boundary. However, the metrics of flawed elements in documents 1 and 4 either remain the same or even increase after correction, suggesting participants have limitations in identifying inaccurate bounding boxes in those two documents. By observing our system's missing and duplicated text output, we found both participants could adjust inaccurate bounding boxes. At the same time, they ignored the creation of new bounding boxes for metadata information in documents. During the analysis, we realized that we did not provide clear instructions on accomplishing the task, leading to participants not being aware of how to capture all the elements in documents. Besides, since participants were given instructions verbally, they had to ask questions to researchers continuously during the research. In addition, participants misunderstood the task of correcting inaccurate bounding boxes. It required participants to ensure bounding boxes covered all corresponding text and then to trim bounding boxes. Sometimes, the bounding box might not accurately align with the intended text. For example, it could cover "by Chenyu Mao" instead of just "Chenyu Mao" for the author label or include "In Partial Fulfillment of the Requirements for the Degree Master of Science" instead of focusing solely on "Master of Science" as the degree label. The bounding box should have precise coverage, so it exclusively contains "Chenyu Mao" for the author label and "Master of Science" for the degree. Accordingly, we added an instruction page with examples in the survey so that participants can refer to it by switching to another browser tab instead of asking for help from researchers, as shown in Appendix C.2.



(a) Participant 1

(b) Participant 2

Figure 6.1: Bar charts illustrating the number of flawed elements across four ETDs, both before and after manual corrections by the Pilot Study Participants.

In most cases, the number of incorrect labels decreases after correction, with the second document showing the most significant improvement, as both participants successfully eliminated all errors. This is presented in Figures 6.2a and 6.2b. However, the number of mistakes in the first document increased after participant corrections, suggesting that participants encountered challenges in identifying the correct labels during the experiment. The exceptional result for the second document can be attributed to our system's failure to capture any metadata from its cover page, which had a novel style. In contrast, it captures partial metadata for documents three and four.

To find what puzzles participants when identifying the wrong labels, we broadly break down the number of wrong labels into several categories, including metadata, figure, algorithm, and paragraph or caption. Since these categories have the most diverse formats in ETDs, our object detection model performs poorly in capturing them. For the two participants, the

breakdown of wrong labels is shown in Tables 6.2 and 6.3, respectively.

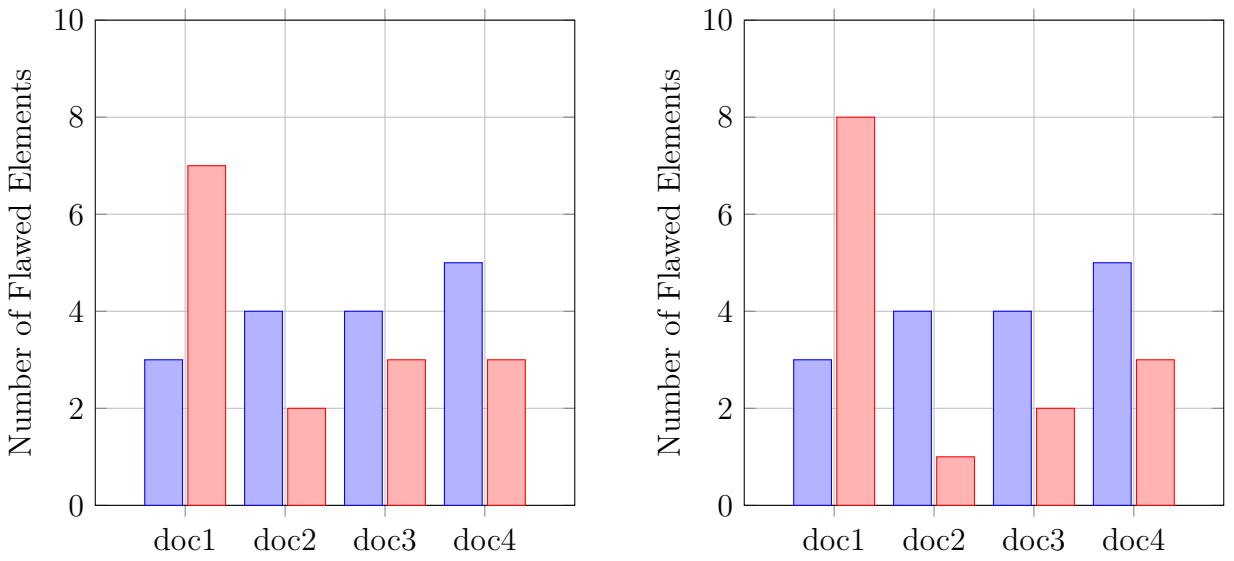
Meanwhile, we also looked into the two chosen experimental ETDs and tried to find what elements confuse participants, especially for Document 1, since the number of flawed elements and wrong labels increased after manual corrections. Our XML template has an independent element representing the thesis abstract, and its content relies on the text extracted by the bounding box labeled as "abstract text". In Document 1, there is a thesis abstract and three chapter abstracts, and both participants labeled some of them as abstract text and some as paragraphs, making the chapter abstract appear in the thesis abstract place. Accordingly, we added the list of element categories to the instruction page and ask users to browse it before they start. Besides we also make the metadata label written in bold text, to get more attention.

Upon analysis of their feedback after the given task, both participants agreed that element annotation without an object detection model would be much more challenging. Participant 1 described the thought of manually identifying and bounding every single element as "daunting." They imagined painstakingly going through each page, looking for different types of elements, and creating bounding boxes from scratch. Participant 2 also endorsed this system, claiming that manual annotation would be highly time-consuming and tend to produce redundant human errors. However, they also pointed out that the process would be significantly smoother if the annotations generated by the system were more accurate.

Table 6.2: Wrong Label Breakdown from Participant 1

doc_idx	**wrong labels**	**metadata**	**figure**	**algorithm**	**caption/paragraph**
1	7	1	0	0	6
2	0	0	0	0	0
3	3	3	0	0	0
4	3	3	0	0	0

Table 6.4 provides a comparative overview of participants based on their performance, aca-



(a) Participant 1 (b) Participant 2

Figure 6.2: Bar charts illustrating the number of wrong and missing labels across four ETDs, both before and after manual corrections by the first group of participants (Group 1).

Table 6.3: Wrong Label Breakdown from Participant 2

doc_idx	wrong labels	metadata	figure	algorithm	caption/paragraph
1	8	1	0	0	7
2	0	0	0	0	0
3	3	3	0	0	0
4	3	3	0	0	0

demic backgrounds, familiarity with ETDs, and average time spent on a document. Although participants were allocated a generous three-hour window to complete the study, our system-logged data revealed that they completed their tasks much more efficiently than anticipated. This unexpectedly quick completion time can be attributed to the initial lack of detailed instructions, as we did not explicitly specify that participants needed to not only correct wrong labels but also create new bounding boxes for missing elements. The system automatically tracked the time spent on each document, measuring from when a participant started working on a single document until they submitted their corrections. The correction

rate was calculated by comparing the number of correctly modified bounding boxes against the total number of known errors in each document, as determined by our ground truth analysis. Both participants come from STEM majors but have different levels of familiarity with ETDs, as Participant 2 is highly familiar with ETDs while Participant 1 is newly introduced to ETDs. The result correlates with their performance, as Participant 2 achieved a higher correction rate in flawed elements and wrong labels than Participant 1. Although Participant 2 performed better in the accuracy of correction work, the average time spent on correction is also higher than Participant 1, indicating a possible trade-off between accuracy and time. Overall, the result reveals that higher familiarity with ETDs improves correction accuracy, and the cost of longer task duration accompanies it.

Table 6.4: Overview of participants' backgrounds and their correction performance, including familiarity levels, correction percentages for flawed elements and wrong labels, and the average time spent (in minutes) on a document.

ID	Major	Familiarity	FE %	WL %	Avg Time (min)
1	STEM	Slightly	53	38	4.75
2	STEM	Extremely	57	47	5.75

6.3 Main User Study

After we completed the pilot study, we gained insight into potential improvements to the formal user study. As suggested by the pilot study results, we applied the following improvement strategies to help formal participants focus on the correction task:

1. Introducing an instruction page between the demographic survey and the post-survey. This allows participants to revisit the instructions as needed without having to seek assistance from researchers.

2. Enhancing the instruction page by including a comprehensive list of all categories, to help users familiarize themselves with the system, thereby reducing the risk of missing elements due to unfamiliarity.

3. Providing detailed guidance on trimming bounding boxes, particularly for metadata labels. For example, participants are instructed to ensure that the edges of the bounding box are tightly aligned with the relevant text, avoiding the inclusion of any unrelated content.

The experimental settings for the main user study were consistent with those outlined in Section 6.1, incorporating additionally only the aforementioned improvements. We recruited eight participants, ensuring a balanced distribution between STEM and non-STEM academic backgrounds, comprising undergraduate students, graduate students, and staff members. Table 6.5 shows the detailed demographic information of these participants, including their gender, academic background, and academic status. For the study materials, we selected an additional twelve ETDs from our pool of ETDs, and established their respective ground truth measurements, as presented in Table 2.2 in Section 2.1. Some participants have already received their payments in accordance with University policies and what was told to the IRB and the other participants. The remaining payments are expected to be completed soon.

Participant ID	**Gender**	**Academic Background**	**Academic Status**
P1	Male	STEM	Master's Student
P2	Female	Non-STEM	Faculty Member
P3	Male	STEM	Undergraduate Student
P4	Female	STEM	Undergraduate Student
P5	Male	Non-STEM	Master's Student
P6	Female	STEM	Ph.D. Student
P7	Male	Non-STEM	Ph.D. Student
P8	Male	Non-STEM	Master's Student

Table 6.5: Demographic Information of Study Participants

6.4 Main User Study Results

Table 6.6 presents the overall data across all ETDs in our study. The first column shows the unique identifier for each ETD document. The second and third columns (FE and WL) represent the initial number of flawed elements and wrong labels output by our system before any human correction. The subsequent columns show the remaining errors after human intervention: columns four and five (FE(P1) and WL(P1)) indicate the number of flawed elements and wrong labels remaining after correction by the first participant in each experimental group, while columns six and seven (FE(P2) and WL(P2)) show the corresponding numbers after correction by the second participant. The participant cohort was strategically divided into four equal groups of two participants each, so each group was representing STEM and non-STEM disciplines, to facilitate comparative analysis. The detailed findings for each group are presented in Figures 6.5 through 6.12, with each group having two figures representing the number of wrong labels and flawed elements before and after human correction.

Across all groups, the general trend is that manual correction reduces the number of flawed elements and wrong labels in most ETDs, which indicates that the manual correction process effectively improves object detection accuracy in the ETDs. For example, in many cases, the height of the bars representing the number of flawed elements and wrong labels after correction is lower than before correction, demonstrating a positive impact of manual intervention. Nevertheless, the specific reduction in flawed elements and wrong labels varies among different ETDs within each group. Some ETDs show a big decrease, while others show a more modest change or even an increase in certain cases. For instance, in Group 1, the reduction in flawed elements and wrong labels is more pronounced in some documents than in others. Figures 6.3 and 6.4 demonstrate the system's performance before and af-

Doc	FE	WL	FE(P1)	WL(P1)	FE(P2)	WL(P2)
1	523	3	3	1	3	0
2	38	4	0	0	0	0
3	15	4	6	0	2	0
4	50	5	50	5	0	3
5	86	4	0	0	0	0
6	321	6	37	1	0	2
7	602	2	2	1	35	1
8	205	14	34	1	2	0
9	317	2	0	1	0	0
10	35	10	3	1	2	0
11	63	8	31	2	2	0
12	29	41	32	21	2	0
13	0	5	0	2	0	2
14	10	2	4	0	2	0
15	1	10	1	0	0	0
16	36	10	1	5	0	1

Table 6.6: Statistical Analysis of Ground Truth Data for ETDs

ter human correction. In Figure 6.3, the object detection model fails to accurately capture several elements and generates imprecise bounding boxes. Figure 6.4 shows the same document after participant corrections, displaying accurate element detection and proper HTML rendering of all document metadata components.

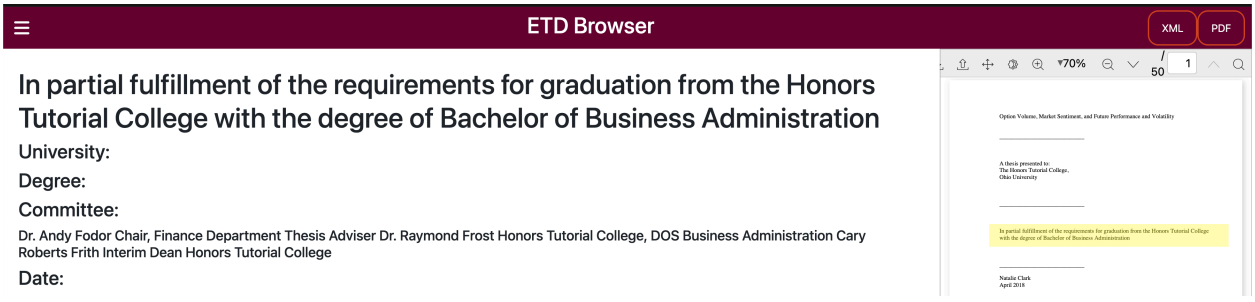


Figure 6.3: ETD browser rendering of model-generated (before human correction) XML output. The model's element detection appears to be inaccurate, as evidenced by incorrect parsing of structural components in the thesis metadata

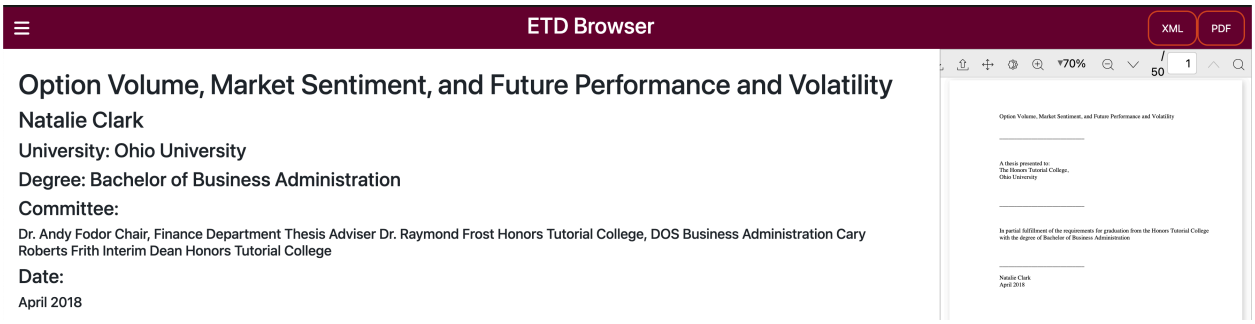
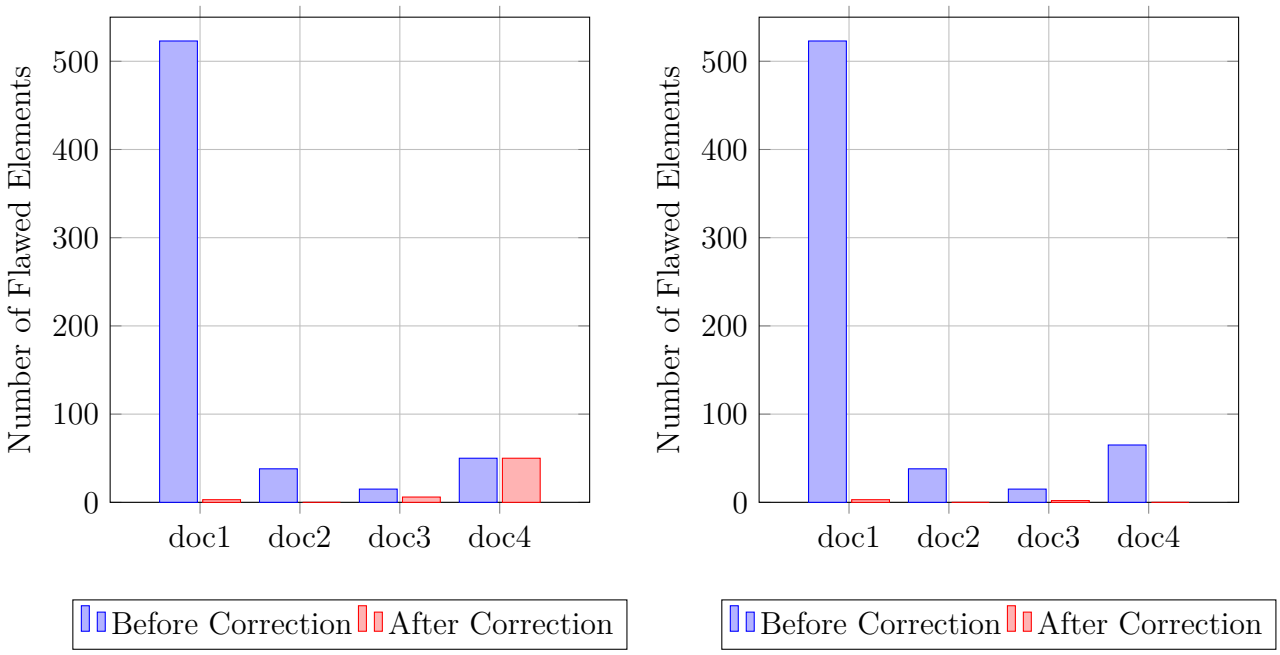


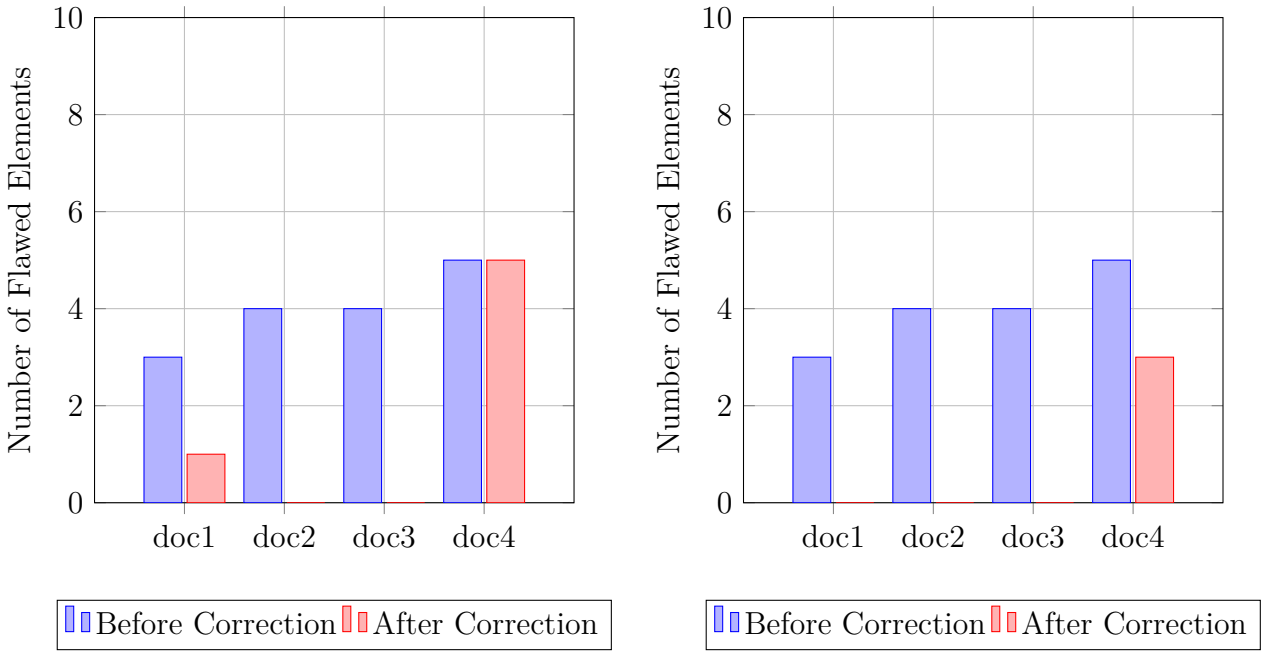
Figure 6.4: ETD browser rendering XML content output after one of the participants corrections, showing accurate detection and structuring of all document metadata elements



(a) Participant 1

(b) Participant 2

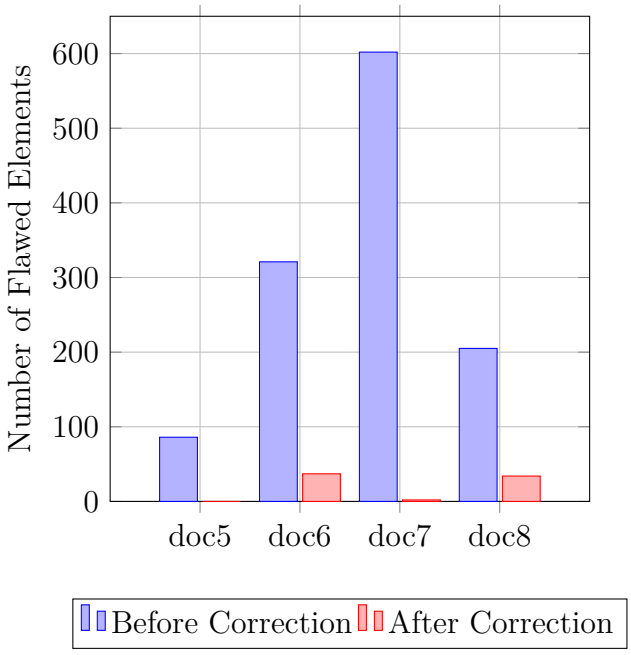
Figure 6.5: Bar charts illustrating the number of flawed elements across four ETDs, both before and after manual corrections by the first group of participants (Group 1).



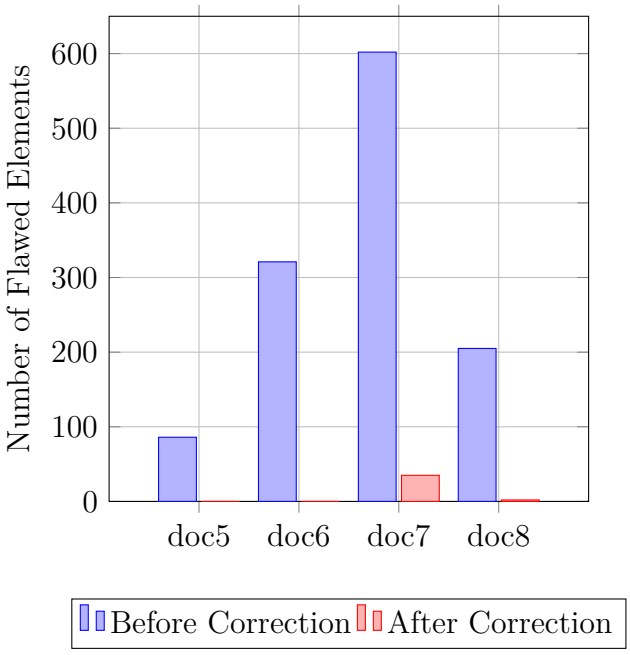
(a) Participant 1

(b) Participant 2

Figure 6.6: Bar charts illustrating the number of wrong and missing labels across four ETDs, both before and after manual corrections by the first group of participants (Group 1).

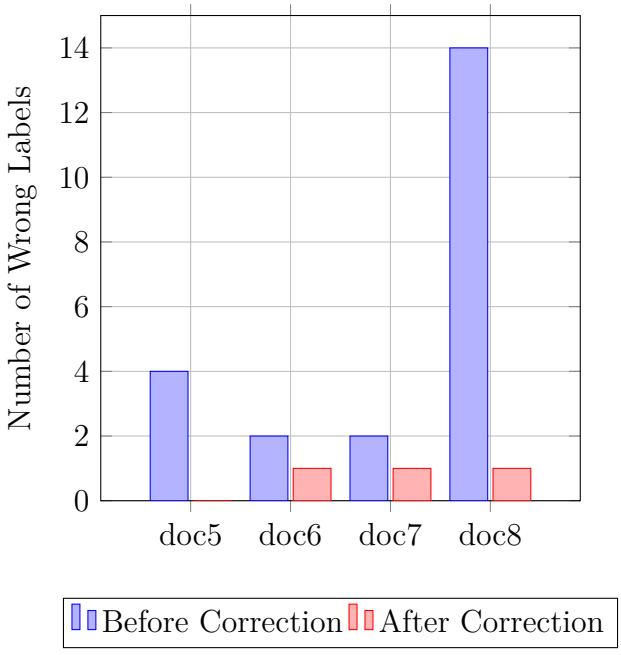


(a) Participant 3

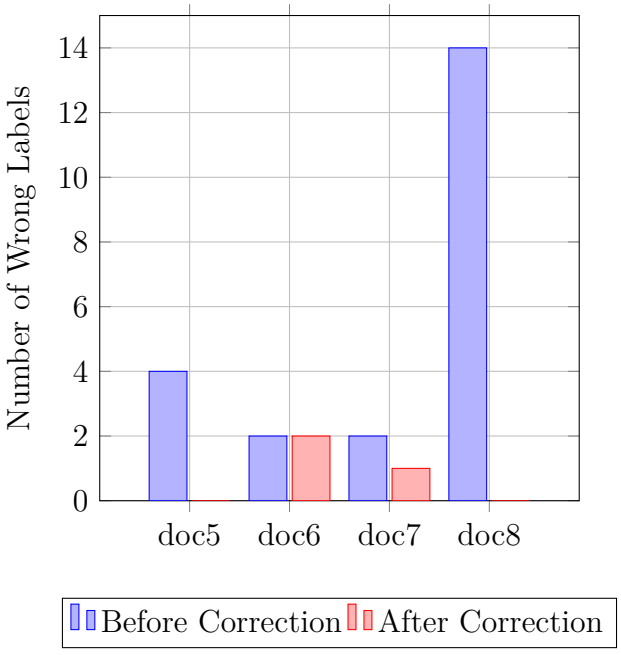


(b) Participant 4

Figure 6.7: Bar charts illustrating the number of flawed elements across four ETDs, both before and after manual corrections by the second group of participants (Group 2).

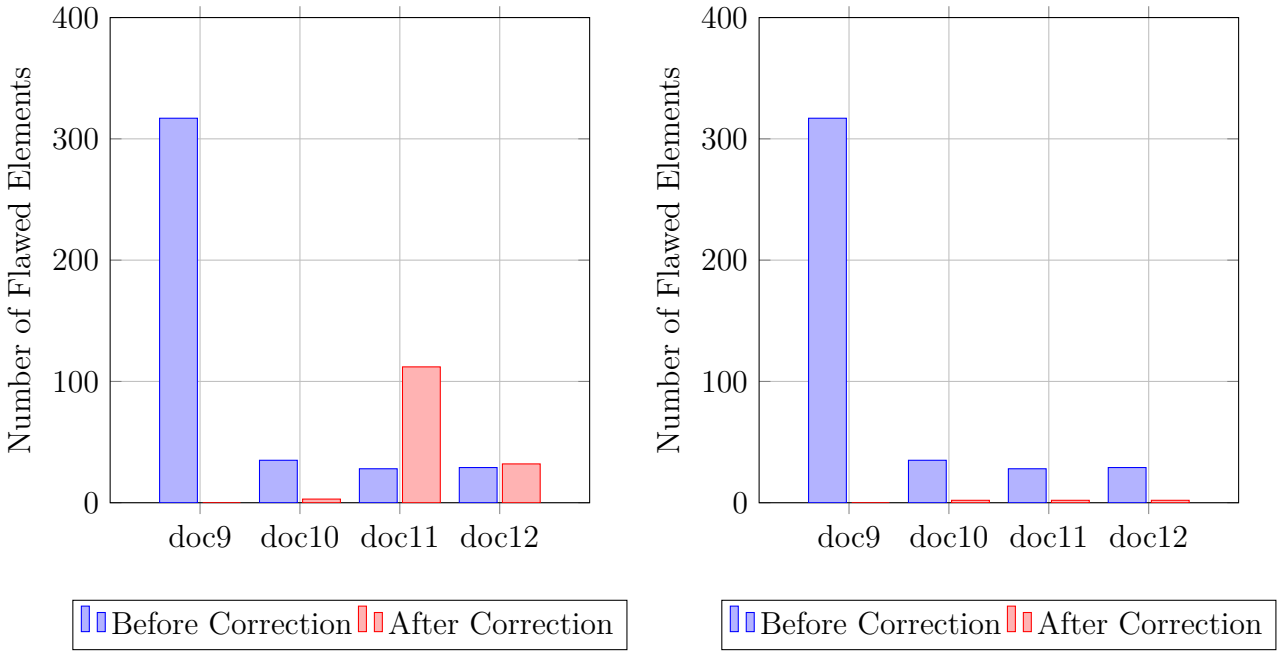


(a) Participant 3



(b) Participant 4

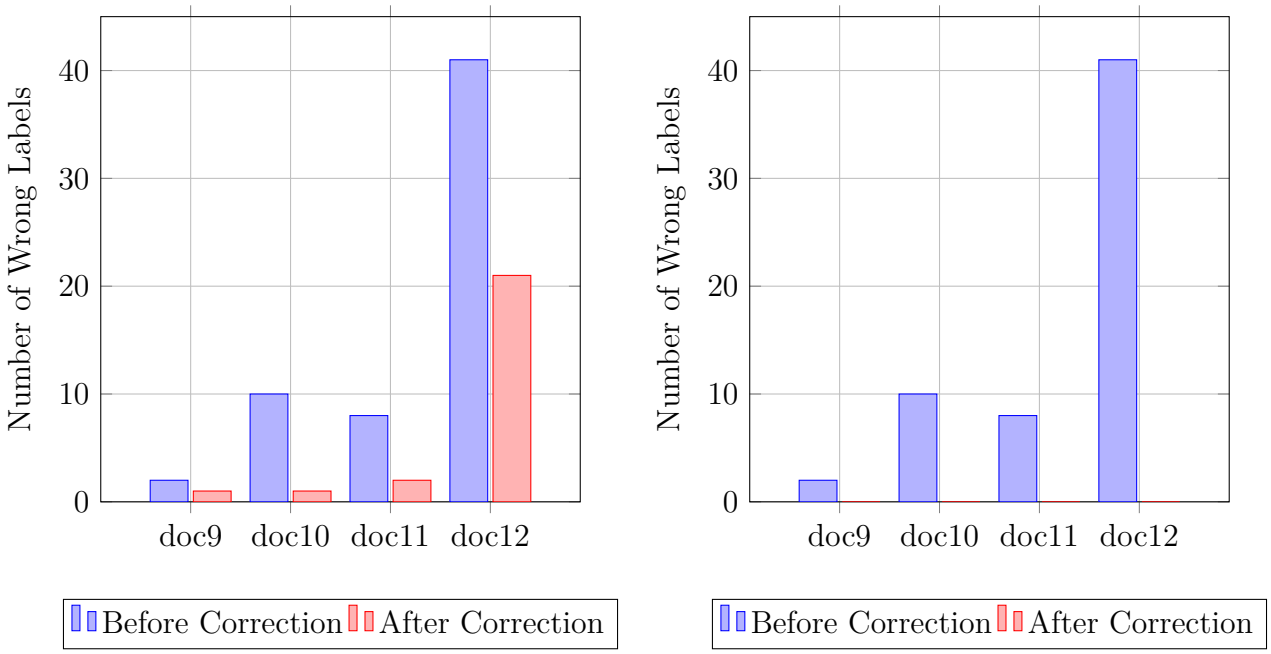
Figure 6.8: Bar charts illustrating the number of wrong and missing labels across four ETDs, both before and after manual corrections by the second group of participants (Group 2).



(a) Participant 5

(b) Participant 6

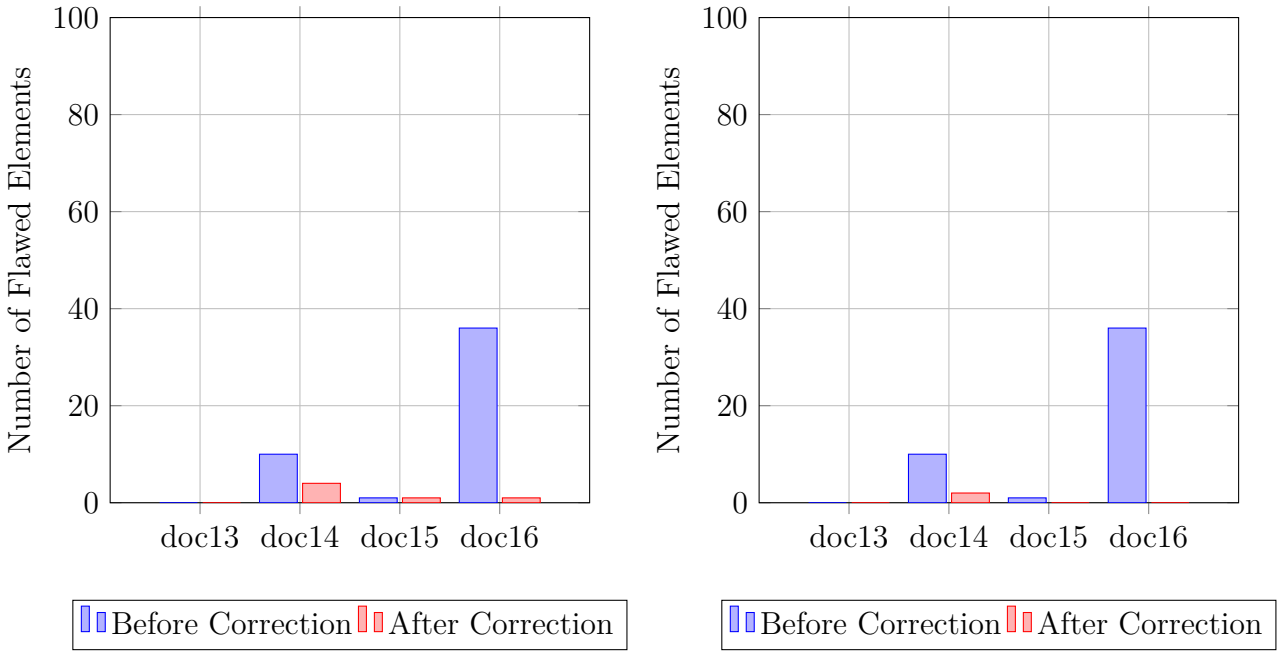
Figure 6.9: Bar charts illustrating the number of flawed elements across four ETDs, both before and after manual corrections by the third group of participants (Group 3).



(a) Participant 5

(b) Participant 6

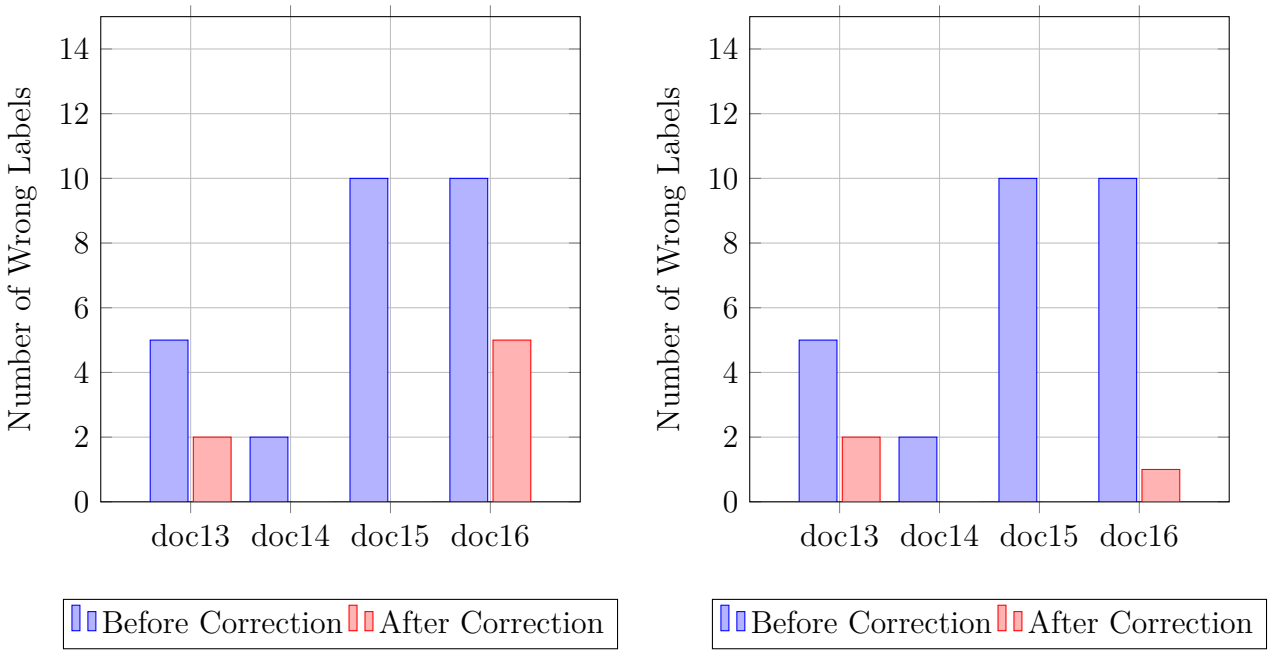
Figure 6.10: Bar charts illustrating the number of wrong and missing labels across four ETDs, both before and after manual corrections by the third group of participants (Group 3).



(a) Participant 7

(b) Participant 8

Figure 6.11: Bar charts illustrating the number of flawed elements across four ETDs, both before and after manual corrections by the fourth group of participants (Group 4).



(a) Participant 7

(b) Participant 8

Figure 6.12: Bar charts illustrating the number of wrong and missing labels across four ETDs, both before and after manual corrections by the fourth group of participants (Group 4).

Chapter 7

Discussion

Table 2.2 reveals potential detection problems across multiple documents. Analysis of our baseline system revealed significant quality issues in element detection and classification. Some Electronic Theses and Dissertations (ETDs) contained over 300 problematic elements. As illustrated in Figure 5.2, one common issue involves incorrect bounding box predictions, where the blue figure box erroneously encompasses both the figure and its caption. Table 2.2 also shows a significant number of misclassified labels, indicating that our object detection model struggles with certain element classifications. For instance, in the same Figure 5.2, the model incorrectly assigns a table label (shown in green) to what is a figure caption, highlighting the challenges our object detection model faces in accurately distinguishing between different document elements. To systematically investigate the underlying elements that contribute to these potential issues, we conducted a detailed analysis of the two main categories: flawed elements and incorrect labels. Statistics for these are shown in Tables 7.1 and 7.2. Flawed content primarily stems from two sources: inaccuracies in textual content and figures. Meanwhile, incorrect labels can be classified into four categories: metadata, figure, algorithm, and caption/paragraph. Among all the errors produced by our system, flawed content accounts for the majority, leading to incomplete and duplicated text appearing in the final XML file. As for the flawed figures, most of them cover not only the figure but also accidentally cover the figure caption and all the text surrounding the figure, which proves that the circumstance that is illustrated in Figure 5.3 is not an isolated case, but

rather a prevalent issue observed across the majority of ETDs. Table 7.2 shows that our object detection model struggles with identifying captions and paragraphs, perhaps because we trained our current model with ETDs from a small range of formats. Thus, it fails to recognize the caption and paragraph occurrences across a larger set of ETDs. Correct handling of captions and paragraphs was only achieved on 4 of the ETDs studied, while in the other ETDs the number of errors with these numbered between 1 and 41.

Table 7.1: Distribution of Flawed Element Types Across ETDs

Doc	**Flawed Element**	**Flawed Text**	**Flawed Figure**
1	523	523	0
2	38	38	0
3	15	15	0
4	50	49	1
5	86	85	1
6	321	316	5
7	602	600	2
8	205	205	0
9	317	317	0
10	35	25	10
11	63	55	8
12	29	29	0
13	0	0	5
14	10	9	1
15	1	0	1
16	36	25	11

Table 7.2: Distribution of Wrong Labels (WL) Types Across ETDs

Doc	WL	Metadata	Figure	Algorithm	Caption/Paragraph
1	3	1	0	0	2
2	4	4	0	0	0
3	4	4	0	0	0
4	5	4	0	0	1
5	4	1	0	2	1
6	6	1	0	0	5
7	2	0	0	0	2
8	14	1	13	0	0
9	2	2	0	0	0
10	10	0	0	0	10
11	8	0	0	0	8
12	41	0	0	0	41
13	5	0	0	0	5
14	2	0	0	0	2
15	10	0	0	0	10
16	10	0	0	0	10

7.1 H1: Manual correction will lead to the identification of many problems with the current system and its object recognition model

The bar charts in Chapter 5 display data in a dual-comparison format, with blue bars showing "Before Correction" and red bars showing "After Correction" across four documents. We can deduce the effectiveness of correction by calculating the height difference. Specifically, we calculated the average correction rate by determining the difference between pre- and post-manual correction values for each group, with the results presented in Table 7.3.

Based on the average performance metrics table, Non-STEM ETDs showed a slightly higher flawed elements correction rate (85.0%) than STEM documents (78.6%), implying that Non-STEM content might be slightly easier to correct for flaws. Nevertheless, the situation

Table 7.3: Comparison of Average Correction Rates Across Documents Before and After Manual Intervention. The N/A (Not Applicable) value is since the model achieved perfect accuracy with zero errors when processing this ETD, making it mathematically impossible to calculate the average correction rate due to a zero denominator in the calculation.

Doc no.	**Type**	**No. pages**	**Avg. FE Correction (%)**	**Avg. WL Correction (%)**
1	Non-STEM	73	99	83
2	Non-STEM	50	100	100
3	STEM	34	73	100
4	STEM	52	50	20
5	STEM	39	100	100
6	Non-STEM	76	94	75
7	Non-STEM	55	97	50
8	STEM	55	91	96
9	Non-STEM	59	100	75
10	STEM	67	93	95
11	STEM	61	73	88
12	Non-STEM	80	41	74
13	STEM	74	N/A	60
14	STEM	42	70	100
15	Non-STEM	48	50	100
16	Non-STEM	74	99	70
STEM Average			78.6	85.6
Non-STEM Average			85.0	78.4
Overall Average			82.0	81.7
Wilcoxon p value			0.00065	0.00005

switches for wrong label correction, where STEM ETDs performed better, with an average of 85.6% compared to Non-STEMs 78.4%. Notably, STEM ETDs show correction rates ranging from at least 50% to 100%, indicating that the complexity or nature of the STEM content might particularly impact the correction judgments. Similarly, Non-STEM ETDs show variation, though generally maintaining higher consistency in Flawed Correction rates (mostly above 90%, with Document 12 being a notable outlier at 41%). The overall averages (82.0% for Flawed Correction and 81.7% for Label Correction) reveal that our human-in-the-loop AI system performs relatively consistently across both metrics, though individual

documents show significant variations. This indicates that while the system is generally effective, specific characteristics of certain ETDs—e.g., novel metadata formats and figures—might affect the correction process regardless of whether they are STEM or non-STEM materials. The Wilcoxon p values were calculated using the signed-rank test which compares sample pairs, with a small sample size. The Wilcoxon signed-rank test first establishes two hypotheses: the null hypothesis (H0) and the alternative hypothesis (H1). H0 suggests that there is no difference between the paired samples, while H1 indicates that there is a difference. A small p-value (less than the significance value) suggests that the observed data is unlikely under the null hypothesis, leading to the rejection of the null hypothesis in favor of the alternative hypothesis. We first took the average number of remaining errors for each ETD and compared them with those before human correction. Subsequently, we employed the Wilcoxon signed-rank test implemented in the Python scipy.stats library to assess statistical significance. The code yielded p-values of 0.00065 and 0.00005 for FE and WL respectively, both of which were substantially lower than the significance threshold (0.05). These results provide strong statistical evidence for significant differences between the compared groups, which strongly supports our hypothesis that manual correction would lead to identifying numerous issues of our baseline system and current object detection model.

7.2 H2: Higher user satisfaction scores will be positively correlated with perceived ease of use

After all participants complete the error correction of ETDs, they are asked to fill out a post-survey that gives ratings to the system and answers some usability questions based on their experience using the system. We adopted a Likert 7-point scale so that users can give rating scores by adjusting the dragging bar, where a minimum score means strongly disagree

and a maximum score means strongly agree. The rating statements are:

1. The system was easy to learn and use.
2. I felt comfortable using the toolbar to correct errors.
3. I was able to complete tasks without needing assistance.
4. I was able to correct most of the errors (e.g., labels or text alignment) without difficulty.
5. I encountered minimal disruptions or technical issues while using the system.

The quantitative Table 7.4 shows high scores across usability metrics, with an average of 5.44 out of 7, while the qualitative feedback provides deeper insights about the given score. To better understand these ratings, we analyzed user responses to two key usability questions: 'How satisfied were you with the tools available for correcting errors? Are there any additional tools or features you would suggest?' and 'Were there any parts of the system that you found confusing or difficult to use? If so, which ones and why?' These questions helped reveal specific usability challenges that influenced the quantitative scores, and their responses are presented in Figures 7.5 and 7.6. Their responses can be broadly divided into two categories, user interface and confusion of elements. Responses show that the current functionality of changing bounding boxes is not easy to adopt, especially when several bonding boxes overlap with each other, since the system requires users to select the bounding boxes before starting to change their coordinates. Moreover, some participants expressed their confusion about the categories, so it would be essential to provide training before users start correction work. The Task Completion Independence showed the most varied responses (ranging from 1 to 7) and one of the lower mean scores, which can be directly linked to several usability challenges identified in the user feedback. Participants reported difficulty when selecting correct labels for titles and subchapters, unclear definitions, and

confusion about the guidelines for the labeling of paragraphs. These issues explain why some users struggled to work independently while others managed well, resulting in a wide score distribution. Technical Issues received the lowest mean score, with ratings ranging from 2 to 6, which aligns with the specific technical challenges reported in the user feedback. Users identified several technical problems, including unexpected pointer tool reloads, the need to refresh pages to make label options appear, and difficulties with zoom functionality. The overlapping tag selection issues and bounding box design problems further contribute to these lower technical ratings.

Table 7.4: Descriptive Statistics of System Usability Dimensions (N=8)

Dimension	**Mean**	**SD**	**Min**	**Max**
System Learnability	5.50	1.20	4.00	7.00
Toolbar Error Correction	5.38	1.30	3.00	7.00
Task Completion Independence	4.88	1.96	1.00	7.00
Error Correction Ease	5.38	1.06	4.00	7.00
Technical Issues	4.63	1.51	2.00	6.00

A Pearson correlation matrix can examine the relationship between two variables, where the correlation coefficient ranges from negative to positive. A positive one (1.0) indicates a perfect positive correlation, and a negative one (-1.0) indicates a perfect negative correlation. In our context, we seek the relationship between user satisfaction scores and perceived ease of use. Among the five rating statements, the fourth statement can be the central factor, which is about error correction ease, given that it directly reflects users' feelings about perceived ease of use. Thus, we calculated the Pearson correlation matrix based on participants' scores and showed interesting correlation patterns with each other dimension, as shown in Table 7.7.

In particular, the result shows that D4 (Error Correction Ease) has strong correlations with both D1 (System Learnability) and D5 (Technical Issues), with r=0.620 and r=0.637, respectively. These numbers demonstrate that the initial system learnability and the technical

Table 7.5: User-reported System Usability Challenges and Pain Points

Question	**User Responses**
Were there any parts of the system that you found confusing or difficult to use? If so, which ones and why?	'Next' button placement too close to 'part 3' button - risk of losing progress if accidentally clicked
	The pointer tool occasionally triggered unexpected reloads
	Difficulty in choosing correct labels for titles, subchapters, and distinguishing between figure titles and captions
	Need to refresh page to make label options appear
	Zoom functionality is difficult to use
	Unclear label definitions - examples needed for confusing categories like 'Chapter subheading'
	Image size adjustment difficulties
	Selection box UI issues - overlapping tags difficult to select

issues strongly influence the perceived ease of use. Nevertheless, D4 (Error Correction Ease) shows weak correlations with both D2 (Toolbar Error Correction) and D3 (Task Completion Independence), with r=0.297 and r=0.163, respectively. These lower correlations are quite interesting as they suggest that the perceived ease of error correction is somewhat independent of both the specific tool usage and users' ability to work autonomously. This unexpected finding indicates that users' perception of ease might be more influenced by the overall system experience (as reflected in technical stability and learnability) rather than specific functional aspects of the error correction process.

To test our second hypothesis, we focused on comparing the measure most directly connected with "perceived ease of use" against each of the other measures in our analysis. These findings challenge our initial assumptions about user satisfaction and perceived ease of use,

Table 7.6: User Satisfaction and Feature Feedback

Question	**User Responses**
How satisfied were you with the tools available for correcting errors? Are there any additional tools or features you would suggest?	7.5/10. Suggested adding explanations of different sections for users less familiar with dissertations/theses
	Need additional labels for 'image' and 'image caption' separate from figure labels. Clearer directions needed regarding individual vs. group paragraph labeling
	Issues with overlapping figure and caption boxes due to photo positioning relative to text
	Generally satisfied
	8/10 satisfaction rating
	8/10. System is easy to use. Suggested adding 'auto match' function to assist with area identification
	Mouse pad needed for precision; mouse cursor settings need adjustment for better precision
	Moderate satisfaction. Bounding box design needs improvement, particularly corner selection

revealing a more complex relationship than anticipated. The implications for future system design are significant: while improving specific tool functionality might seem intuitive, our data suggests that prioritizing system stability, initial learnability, and cognitive accessibility would be more effective in enhancing user experience. Therefore, development efforts should focus on building system robustness and making existing features more accessible rather than adding new functionalities.

Table 7.7: Pearson Correlation Matrix with Focus on Error Correction Ease (D4)

Dimension	D4	D1	D2	D3	D5
D4. Error Correction Ease	1.000				
D1. System Learnability	0.620*	1.000			
D2. Toolbar Error Correction	0.297	0.872**	1.000		
D3. Task Completion Independence	0.163	0.823**	0.805**	1.000	
D5. Technical Issues	0.637*	0.754**	0.519*	0.708**	1.000

Note: * $p < 0.05$, ** $p < 0.01$

Table 7.8: Distribution of Ratings Across Dimensions

Dimension	1	2	3	4	5	6	7	Total
System Learnability	0	0	0	2	2	3	1	8
Toolbar Error Correction	0	0	1	1	2	3	1	8
Task Completion Independence	1	0	0	2	2	1	2	8
Error Correction Ease	0	0	0	2	2	3	1	8
Technical Issues	0	1	1	1	2	3	0	8

7.3 H3: Users with more significant academic experience and domain-specific knowledge will perform better at our task

We examined user performance based on two key factors: academic background (STEM vs. Non-STEM) and familiarity with ETDs.

Since domain-specific knowledge is difficult to assess without an analysis of a large number of domains/disciplines and a correspondingly large number of participants in an experiment, we simplified by grouping domains just into STEM and non-STEM, assuming that STEM participants would have domain knowledge about STEM ETDs, and non-STEM participants would have domain knowledge about non-STEM ETDs. We built a table presenting the average time and correction rate for participants with each of the two academic backgrounds;

see Table 7.9.

In our experiment, we use the level of familiarity with ETDs as the measurement of academic experience, where level 4 represents extremely familiar and level 1 represents unfamiliar. Due to our targeted recruitment process focusing on faculty members and students with ETD experience, participants with Level 1 familiarity (indicating minimal ETD knowledge) were naturally filtered out. This deliberate selection criterion ensured that all participants had at least a basic working knowledge of ETDs. Participants with higher levels of familiarity (3 and 4) demonstrated better correction capabilities than those with level 2 familiarity. Specifically, participants with the highest level of familiarity achieved the highest precision in correcting incorrect labels (87.50%), while participants with the second highest level of familiarity performed best in corrected text (85.50%). These results reveal that users with a rich experience with ETDs can demonstrate an enhanced accuracy of correction.

However, the relationship between academic background and performance gives us unexpected results. Notably, non-STEM participants consistently exceeded their STEM peers in both correction tasks, contrary to our initial assumption since we believed that STEM users are experienced with figures, tables, and algorithms. It turns out that non-STEM participants achieved significantly higher accuracy rates in flawed content correction (96.05% vs. 67.03%) and wrong label correction (90.21% vs. 72.91%). Moreover, non-STEM participants completed their tasks more quickly than STEM participants, with average processing times of 15.31 minutes and 23.19 minutes, respectively.

Participants with familiarity level 4 achieved relatively high performance, suggesting that success in document correction tasks may depend more on general text-processing abilities than on domain-specific technical knowledge. The outstanding performance of non-STEM participants might be attributed to the following factors. First, non-STEM users are likely to develop stronger general proofreading skills than STEM users, since non-STEM users

are more experienced in document processing and textual editing tasks. This long-term exposure to textual content makes non-STEM users more sensitive to errors and quicker at identifying inaccuracies. Second, STEM and non-STEM users exhibit different attention distribution patterns. While STEM users focus on logical and technical details, non-STEM users prioritize content and form. These distinct reading habits enable non-STEM users to detect a broader range of document errors more effectively.

The positive correlation between academic experience and performance aligns with our expectations, indicating that increased experience with ETDs does enhance correction effectiveness and efficiency. However, the stark contrast between STEM and non-STEM performance suggests that the ability required for effective error correction may relate more to general text processing abilities rather than technical domain knowledge.

Table 7.9: Performance Analysis by Major Area

Major Area	**Avg Time**	**Avg FE(%)**	**Avg WL(%)**
STEM	23.19	67.03	72.91
Non-STEM	15.31	96.05	90.21

Table 7.10: Performance Analysis by ETDs Familiarity Level

Familiarity Level	**Count**	**Avg FE(%)**	**Avg WL(%)**
Level 2	2	75.43	73.75
Level 3	3	85.50	80.83
Level 4	3	81.66	87.50

Chapter 8

Contributions and Future Work

8.1 Contributions

This research explored the effectiveness and efficiency of manual correction strategies in improving object detection accuracy for document element recognition systems, particularly for ETDs. Through a user study involving 8 participants working with and evaluating a web-based application, the research investigated how individuals with diverse backgrounds interact with a human-in-the-loop AI system for detecting, parsing, and correcting document content. We found that manual correction significantly improved document recognition accuracy, with non-STEM participants achieving higher correction rates (96.05% for flawed content, 90.21% for wrong labels) compared to STEM participants (67.03% and 72.91%, respectively). Meanwhile, we also evaluated the usability of our system based on the rating score provided by participants. Notably, user satisfaction was strongly correlated with system learnability and stability rather than specific tool functionality, suggesting that overall system experience plays a crucial role in perceived ease of use. Since participants have diverse academic backgrounds that might influence the correction effectiveness and efficiency, we found that higher familiarity with ETDs positively impacted correction accuracy, with the most experienced users achieving 87.50% accuracy in label correction.

The research developed a novel methodology that allows users to correct flawed elements generated by the object detection model in real time. These corrections are incorporated

into the training corpus for continuous model improvement. The system demonstrated effectiveness in addressing common challenges in document recognition, particularly in handling diverse document formats and specialized content across different academic disciplines.

This work contributes to improving document accessibility by establishing effective strategies for human-machine collaboration in document recognition systems, while providing insights for future interface design and workflow optimization in manual correction activities.

8.2 Future Work

Although we conducted this user study to test the effectiveness of our object detection system, several difficulties can limit generalizing the conclusions of our evaluation. Firstly, we only recruited eight participants for this user study, even though we tried to reach all kinds of departments across the university so as to ensure generalizability. Secondly, due to the complicated functionalities provided by our system, since participants started their work without any professional training but only with a simple demonstration and an oral tutorial, their unfamiliarity with using this system might have influenced their performance on identification and correction tasks.

In examining our second hypothesis, we limited our analysis to comparing the measure most directly associated with "perceived ease of use" against other individual measures. While this approach provided basic insights, a more comprehensive evaluation of the relationship between perceived ease of use and user satisfaction would be more convincing from implementing the System Usability Scale (SUS), an industry standard survey for usability assessment [4]. The SUS criteria would offer a more robust framework for evaluating user experience, as it covers multiple dimensions of usability with questions from ten dimensions.

As for our third hypothesis, while our current study provides interesting preliminary insights into the relationship between academic background and correction performance, a more rigorous investigation is needed to validate these findings. Future research should employ a larger sample size ($n > 30$) to enable the use of Analysis of Variance (ANOVA). This powerful statistical method can determine whether there are statistically significant differences between means of multiple groups while controlling for various factors. ANOVA would be particularly valuable in this context as it could help isolate the effects of academic background from other variables that might influence performance, such as familiarity with ETDs.

Bibliography

[1] Aman Ahuja, Alan Devera, and Edward Alan Fox. Parsing Electronic Theses and Dissertations Using Object Detection. In Tirthankar Ghosal, Sergi Blanco-Cuaresma, Alberto Accomazzi, Robert M. Patton, Felix Grezes, and Thomas Allen, editors, *Proceedings of the First Workshop on Information Extraction from Scientific Publications (WIESP 2022), held in conjunction with AACL-IJCNLP 2022*, pages 121–130, Taipei and Online, November 2022. Association for Computational Linguistics. URL: https://ui.adsabs.harvard.edu/WIESP/2022/Schedule.

[2] Aman Ahuja, Kevin Dinh, Brian Dinh, William A. Ingram, and Edward A. Fox. A New Annotation Method and Dataset for Layout Analysis of Long Documents. In *Proceedings of the 3rd International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment (Sci-K 2023)*, Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), pages 834–842, New York, NY, USA, April 30 - May 4 2023. Association for Computing Machinery.

[3] Eugene Belval. pdf2image. https://github.com/Belval/pdf2image, 2024. Python PDF to Image Conversion Library [Accessed: December 12, 2024].

[4] John Brooke. SUS: A Quick and Dirty Usability Scale. In Patrick W. Jordan, Bruce Thomas, Ian Lyall McClelland, and Bernard Weerdmeester, editors, *Usability Evaluation in Industry*, chapter 21, pages 189–194. Taylor & Francis, London, 1996.

[5] Chengliang Chai, Lei Cao, Guoliang Li, Jian Li, Yuyu Luo, and Samuel Madden. Human-in-the-loop Outlier Detection. In *Proceedings of the 2020 ACM SIGMOD In-*

ternational Conference on Management of Data, SIGMOD '20, pages 19–33, New York, NY, USA, 2020. Association for Computing Machinery.

[6] Chengliang Chai and Guoliang Li. Human-in-the-loop Techniques in Machine Learning. *IEEE Data Eng. Bull.*, 43(3):37–52, 2020.

[7] Richeng Cheng. A survey: Comparison between Convolutional Neural Network and YOLO in image identification. *Journal of Physics: Conference Series*, 1453:012139, 2020. 2nd International Conference on Computer Information Science and Artificial Intelligence (CISAI 2019), Xi'an, China, 25-27 October 2019.

[8] Brad Dwyer, Joseph Nelson, Taylor Hansen, Jacob Taylor, and Brad Miller. Roboflow: Computer Vision Infrastructure for Developers. https://roboflow.com, 2024. A comprehensive platform for computer vision development, including tools for dataset management, model training, and deployment [Accessed: December 6, 2024].

[9] Matthias Lee, Samuel Hoffstaetter, Heiki Jauhiainen, Juarez Patel, et al. Python-Tesseract: A Python wrapper for Google's Tesseract-OCR. https://github.com/madmaze/python-tesseract, 2007. An optical character recognition (OCR) tool that provides Python bindings for Google's Tesseract-OCR Engine [Accessed: December 1, 2024].

[10] Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. Document Layout Analysis with an Enhanced Object Detector. In *2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pages 1–5, 2021.

[11] Pallets Team, Armin Ronacher, David Lord, Adrian Mönnich, Philip McKay, et al. Flask: A Lightweight WSGI Web Application Framework. https://flask.palletsprojects.com/, February 2024. [Accessed: December 20, 2024].

[12] QuestionPro Inc. QuestionPro: Online Survey Software for Research & Data Collection. https://www.questionpro.com/academic/39.html, 2005. [Accessed: December 1, 2024].

[13] Jörg Raithel, Ruikai Nettekoven, Jorj X. Liu, et al. PyMuPDF (fitz): Python bindings for the MuPDF PDF toolkit. https://github.com/pymupdf/PyMuPDF, 2016. A high-performance PDF, XPS, and E-book processing library with Python bindings [Accessed: December 7, 2024].

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. https://arxiv.org/abs/1506.02640, 2016. A groundbreaking paper introducing YOLO, a state-of-the-art real-time object detection system [Accessed: October 29, 2024].

[15] Nina Rybárová. Python-Flask for creating Simple Web Applications. *Journal of Information Technology and Applications*, 11(2):65–72, December 2022. Special issue on Web Technologies and Frameworks.

[16] Tehreem Shehzadi, Didier Stricker, and Muhammad Zeshan Afzal. A Hybrid Approach for Document Layout Analysis in Document Images. *Lecture Notes in Computer Science*, 14474:21–39, 2024. Part of the 16th IAPR International Workshop on Document Analysis Systems (DAS 2024).

[17] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, July 2022. Code available at https://github.com/WongKinYiu/yolov7 [Accessed: December 20, 2024].

[18] Severin Wörner and React Image Annotate Contributors. React-Image-Annotate: An Open Source Image Annotation Library. https://github.com/waoai/react-image-annotate, March 2024. [Accessed: December 1, 2024].

[19] Qintong Zhang, Victor Shea-Jay Huang, Bin Wang, Junyuan Zhang, Zhengren Wang, Hao Liang, Shawn Wang, Matthieu Lin, Wentao Zhang, and Conghui He. Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction. *arXiv preprint arXiv:2410.21169*, 2024.

Appendices

Appendix A

Consent

Information Sheet for Participation in a Research Study

Principal Investigator: Dr. Edward A. Fox, (540) 231-5113 or fox@vt.edu

IRB# and Title of Study: IRB # 24-1189 Exploring the Role of Manual Correction in Document Recognition: User Study on Performance and Usability

Sponsor: IMLS

You are invited to participate in a research study. This form includes information about the study and contact information if you have any questions. WHAT SHOULD I KNOW?

Purpose of the Study: The purpose of this study is to explore and compare how different individuals interact with a human-in-the-loop AI system designed for detecting, parsing, and correcting information from electronic theses and dissertations (ETDs).

Background: Computer Vision: Computer vision is a field of artificial intelligence that enables machines to interpret and process visual information from the world, such as images or videos, to perform tasks like recognition, detection, and analysis. Object Detection: Object detection is a computer vision task that identifies and locates objects within images or videos by classifying them and marking their positions with bounding boxes.

Study Process: You must come in person to 2030 Torgersen Hall and use the application setup on a lab computer to complete the study. Your task is to use our homegrown system

to correct the wrong labels and inaccurate bounding boxes in an XML file, which will be displayed in a rendered interactive HTML page that our computer software generates after trying to identify the various document elements (e.g., paragraph, equation, page number). We will not collect user data other than when using our homegrown web-based system, which will be set up for you so you don't need to log in or provide any personal information aside from general demographics. During this process, there will not be any video or audio capture. However, our system will record all user actions related to our system as you complete the task; we ask you to stay focused on the given task. In addition, please do not open any other program or application on our computer during the process.

1. First, you will be guided to the VT version of the QuestionPro platform to complete a survey where you should carefully read the background information about this research. You can move forward only after you give consent.

2. After providing consent through QuestionPro, you will be asked to answer some demographic questions, such as age, major, and category (undergraduate or graduate student, member of faculty or staff, researcher). If you meet the specified eligibility criteria, you will receive detailed instructions on using the system.

3. Next, you will be given 4 ETDs and use the toolbar, mouse, and keyboard to correct the errors (wrong labels or inaccurate bounding boxes) occurring in the XML files, and then have the corrected XML result saved by clicking the "save" button on the toolbar. Though focused on high quality work, you should try to complete the checking of all 4 ETDs in the 3 hours allowed. It also is fine if you complete the checking in less than 3 hours.

4. Finally, after you finish the annotation of as much as you can in the given set of ETDs, you will be asked to return to the QuestionPro form to provide feedback as

well as ratings based on standard usability evaluation criteria (such as ease of use, satisfaction, and learnability) using a Likert scale (e.g., 1 = Strongly Disagree to 7 = Strongly Agree).

You will be paid for up to 3 hours at the rate of $14/hour, prorated to account for the time spent in the experiment. The payment does not depend on the quality of the work, though we ask you to make a best effort. We will collect your name and contact information so payment can be made. If you decide to participate in this study, you will complete a survey. As part of the study, you will follow the procedures above. The study should take approximately 3 hours of your time. We do not anticipate any risks from completing this study.

If you agree to take part in this research study, you will receive $14 per hour, prorated for your time working on the task.

You can choose whether to be in this study or not. If you volunteer to be in this study, you may withdraw at any time without consequences of any kind. You may also refuse to answer any questions you don't want to answer and remain in the study. The investigator may withdraw you from this research if circumstances arise which warrant doing so. Participation will be entirely voluntary. If you are a student, your decision to participate, and participation, will have no impact on your academic standing, grades, or relationships with faculty. If you are a researcher or member of the faculty or staff, your decision to participate, and participation, will have no impact on your work status. You can withdraw from the study without penalty.

CONFIDENTIALITY

We will do our best to protect the confidentiality of the information we gather from you, but we cannot guarantee 100% confidentiality. Any data collected during this research

study will be kept confidential by the researchers. Your personal information (e.g., name and contact information, used for making payment) will be securely stored separately from your responses. The collected information will be uploaded to a secure, password-protected storage space accessed only by the research team. The survey is hosted on a secure platform approved by Virginia Tech. All data will be stored and handled in compliance with data protection regulations. The information will be stored for 5 years after the study has been completed and then destroyed.

WHO CAN I TALK TO? If you have any questions or concerns about the research, please feel free to contact Chenyu Mao (mchenyu@vt.edu). You are not waiving any legal claims, rights, or remedies because of your participation in this research study. If you have questions regarding your rights as a research participant, contact the Virginia Tech HRPP Office at 540-231-3732 (irb@vt.edu). By clicking "I Consent," you confirm that you have read and understood this information and voluntarily agree to participate in the study.

Appendix B

Recruitment Method

Email recruitment: Recruitment messages will be sent to potential participants via university-approved mailing lists and direct emails. These emails will include a brief description of the study, eligibility criteria, participation instructions, and a statement emphasizing participation is voluntary and individuals should not feel pressured to join the study, even if the recruitment material is shared by someone we directly contacted. For follow-up communications, we will limit contact to a maximum of one initial email and no more than two follow-up emails.

Subject: Invitation to Participate in Virginia Tech Research Study on human-in-the-loop AI system

Email Content: Dear Students, Faculty, and Researchers, We are conducting a research study at Virginia Tech focused on exploring the role of manual correction in document recognition, for a variety of participants. Your insights as someone with experience in academic content, particularly electronic theses and dissertations (ETDs), will be valuable for this research.

Study Details: Eligibility Criteria: Current Virginia Tech graduate students, undergraduate students with research experience, faculty members, or researchers. Purpose: The purpose of this study is to explore and compare how different individuals interact with a human-in-the-loop AI system designed for detecting, parsing, and correcting information

from ETDs. Time Commitment: Approximately 3 hours.

Procedures: You must come in person to 2030 Torgersen Hall and use the application setup on a lab computer to complete the study. Your task is to use our homegrown system to correct the wrong labels and inaccurate bounding boxes in an XML file, which will be displayed in a rendered interactive HTML page that our computer software generates after trying to identify the various document elements (e.g., paragraph, equation, page number). We will not collect user data other than when using our homegrown web-based system, which will be set up for you so you don't need to log in or provide any personal information aside from general demographics. During this process, there will not be any video or audio capture. However, our system will record all user actions related to our system as you complete the task; we ask you to stay focused on the given task. In addition, please do not open any other program or application on our computer during the process.

1. First, you will be guided to the VT version of the QuestionPro platform to complete a survey where you should carefully read the background information about this research. You can move forward only after you give consent.

2. After providing consent through QuestionPro, you will be asked to answer some demographic questions, such as age, major, and category (undergraduate or graduate student, member of faculty or staff, researcher). If you meet the specified eligibility criteria, you will receive detailed instructions on using the system.

3. Next, you will be given 4 ETDs and use the toolbar, mouse, and keyboard to correct the errors (wrong labels or inaccurate bounding boxes) occurring in the XML files, and then have the corrected XML result saved by clicking the "save" button on the toolbar. Though focused on high quality work, you should try to complete the checking of all 4 ETDs in the 3 hours allowed. It is also fine if you complete the check in less than 3

hours.

4. Finally, after you finish the checking/annotation of as much as you can in the given set of ETDs, you will be asked to return to the QuestionPro form to provide feedback as well as ratings based on standard usability evaluation criteria (such as ease of use, satisfaction, and learnability) using a Likert scale (e.g., 1 = Strongly Disagree to 7 = Strongly Agree).

You will be paid for up to 3 hours at the rate of $14/hour, prorated to account for the time spent in the experiment. The payment does not depend on the quality of the work, though we ask you to make a best effort. If you are a student, your decision to participate, and participation, will have no impact on your academic standing, grades, or relationships with faculty. If you are a researcher or member of the faculty or staff, your decision to participate, and participation, will have no impact on your work status. You can withdraw from the study without penalty. We will collect your name and contact information so payment can be made.

Contact Information: For more information about the study, or if you have any questions, please contact the research team at mchenyu@vt.edu.

https://virginiatech.questionpro.com/t/Abgn8Z4cWX

Best regards,

Chenyu Mao

Virginia Tech | Computer Science

Blacksburg, VA 24061

540-557-8625 | mchenyu@vt.edu

Appendix C

User Survey

C.1 Demographic Survey

1. Name:
2. Email:
3. Major:

Exercise 1.

Are you 18 years of age or older?

1. Yes
2. No

Exercise 2.

What is your academic status?

1. Undergraduate
2. Graduate
3. Faculty

4. Researcher

5. NA

Exercise 3.

How familiar are you with academic content such as theses or dissertations?

1. Not at all familiar

2. Slightly familiar

3. Moderately familiar

4. Very familiar

5. Extremely familiar

C.2 Instructions

Now, please go back to the web system opened in another tab.

In this section, you will correct the wrong labels and inaccurate bounding boxes in a set of ETDs, and you need to upload a single PDF, annotate it, and save the result. You must repeat this procedure individually until you finish all ETDs in the directory.

1. The first page of the framework provides a straightforward interface with two buttons for file uploading. Since we are annotating four documents, please choose the batch upload button. Afterward, the system will ask you to browse the file and select a document to upload. After you select the document, please wait until the progress bar reaches the end and shows an uploading successful animation, which might take a few seconds. Finally, click the upload button to upload the document to the system.

2. After that, you will see a browsing tool containing the first page of the uploaded document with different bounding boxes and labels; each bounding box represents an element, such as a paragraph, title, and figure. Correct the wrong labels and inaccurate bounding boxes with the provided toolbar. Sometimes, the system might fail to capture some elements on that page; please use the toolbar to create a new bounding box to cover that. The first mouse icon enables users to select a single bounding box. Then, a selection box will pop up, allowing the user to browse and correct the most suitable category option with a drop-down button. Besides, when the mouse hovers over the four corners of the bounding box, it transforms into a drag handle, allowing the user to adjust its size. The second-hand icon enables users to move the position of the document and makes it easier to make corrections. In addition, the third magnifying glass button allows users to zoom in and out of the document, more precisely adjusting the bounding box coverage. Last but not least, the fifth bounding box icon can create a new bounding box, which can be used to capture elements missed by the model inference. The top toolbar is served for page browsing; the previous and next buttons let the user move to the last and next page, while the full-screen button can turn the browser into full-screen mode.

3. If you click the next button and nothing happens, it means you have finished verifying and correcting every page in the document, and you need to save the result by clicking the save button.

4. Change the web address to http://127.0.0.1:5000/reset and work on the next file.

5. After you have done all ETDs, please return to the survey and fill in the rest of the questions.

Below are the categories of all elements in ETDs, please try to correct the bounding box to

the most suitable one and capture the missing label(especially the categories in bold text).

Table C.1: ETD Elements

Metadata
Chapter Title
Chapter Subheading
Title
Abstract Heading
Abstract Text
Algorithm
Author
Committee
Date
Degree
Equation
Equation Number
Figure
Figure Caption
Foot Note
List of Content Heading
List of Content Text
Page Number
Paragraph
Reference Heading
Reference Text
Table
Table Caption
University

Tips:

1. Sometimes, the bounding box might not accurately align with the intended text. For example, it could cover “by Chenyu Mao” instead of just “Chenyu Mao” for the author label, or include “object detection by Chenyu Mao” instead of focusing solely on “object detection” as the title. Please adjust the bounding box to ensure precise coverage, so it exclusively contains “Chenyu Mao” for the author label and “object detection” for the title.

2. Please ensure that when adjusting a bounding box, its edges are tightly aligned with the text without covering any unrelated text.

3. Please ensure that the bounding boxes are not overlapping.

4. If a single bounding box can cover a piece of text, avoid using multiple bounding boxes.

5. After reviewing the instructions above, please ask the researcher to give you a demo of how to use the system.

C.3 Usability Survey

Please rate the following statements based on your experience using the system. Select the response that best reflects your level of agreement (1 = Strongly Disagree, 7 = Strongly Agree).

1. The system was easy to learn and use.

2. I felt comfortable using the toolbar to correct errors.

3. I was able to complete tasks without needing assistance.

4. I was able to correct most of the errors (e.g., labels or text alignment) without difficulty.

5. I encountered minimal disruptions or technical issues while using the system.

Please answer the following questions.

1. Were there any parts of the system that you found confusing or difficult to use? If so, which ones and why?

2. How satisfied were you with the tools available for correcting errors? Are there any additional tools or features you would suggest?

Appendix D

IRB Approval Letter

Division of Scholarly Integrity and Research Compliance
Institutional Review Board
North End Center, Suite 4120 (MC 0497)
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-3732
irb@vt.edu
http://www.research.vt.edu/sirc/hrpp

MEMORANDUM

DATE: November 19, 2024

TO: Edward Fox, Chenyu Mao

FROM: Virginia Tech Institutional Review Board (FWA00000572)

PROTOCOL TITLE: Exploring the Role of Manual Correction in Document Recognition: User Study on Performance and Usability

IRB NUMBER: **24-1189**

Effective November 19, 2024, the Virginia Tech Human Research Protection Program (HRPP) determined that this protocol meets the criteria for exemption from IRB review under 45 CFR 46.104 (d) category(ies) 2(ii),3(i)(B).

Ongoing IRB review and approval by this organization is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities impact the exempt determination, please submit an amendment to the HRPP for a determination.

This exempt determination does not apply to any collaborating institution(s). The Virginia Tech HRPP and IRB cannot provide an exemption that overrides the jurisdiction of a local IRB or other institutional mechanism for determining exemptions.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

https://secure.research.vt.edu/external/irb/responsibilities.htm

(Please review responsibilities before beginning your research.)

PROTOCOL INFORMATION:

Determined As:	**Exempt, under 45 CFR 46.104(d) category(ies) 2(ii),3(i)(B)**
Protocol Determination Date:	**November 8, 2024**

ASSOCIATED FUNDING:

The table on the following page indicates whether grant proposals are related to this protocol.

SPECIAL INSTRUCTIONS:

This amendment, authorized on November 19, 2025, alters the intent of the study to where it is now generalizable.

Date*	OSP Number	Sponsor

* Date this proposal number was added.

If this protocol is to cover any other grant proposals, please contact the HRPP office (irb@vt.edu).