**A**

**PROJECT SCHOOL REPORT**

**ON**

# RE-DACT: AUTOMATED REDACTION

**Submitted By**

**GUNTU CHARITASRI 245522733153**

**THOTA YASHWANTH 245522748056**

**BHEEMANPALLY ABHIJITH 245522748073**

**PANKAJ DESHMUKH 245522748075**

**ODAPALLI KEERTHANA 245522748106**

**TADIVAKA HASMITA 245522748122**

**Under the guidance**

**of**

**Dr. Shilpa Choudhary**

**Assistant Professor CSE (AIML)**



# KESHAV MEMORIAL ENGINEERING COLLEGE

Kachavanisingaram Village, Hyderabad, Telangana 500058.

**January, 2025**

**KESHAV MEMORIAL ENGINEERING COLLEGE**

A Unit of Keshav Memorial Technical Education (KMTES)

Approved by AICTE, New Delhi & Affiliated to Osmania University, Hyderabad

# CERTIFICATE

*This is to certify that the project work entitled* “**RE-DACT: AUTOMATED REDACTION ”** *is a bonafide work carried out by* **“GUNTU CHARITASRI, THOTA YASHWANTH, BHEEMANPALLY ABHIJITH, PANKAJ DESHMUKH, ODAPALLI KEERTHANA, TADIVAKA HASMITA”** of III-year V semester **Bachelor of Engineering** *in* **CSE/ CSE (AIML)** *during the academic year* **2024-2025 and** *is a record of bonafide work carried out by them*.

## Project Mentor

Dr.Shilpa Choudhary

Assistant Professor CSE(AIML)

**ABSTRACT**

In the digital era, safeguarding sensitive information in documents is imperative to prevent unauthorized access and ensure compliance with privacy regulations. To address this challenge, we developed an advanced automated redaction system designed to identify and securely redact Personally Identifiable Information (PII). The system leverages Tesseract OCR to extract text from diverse document formats, including PDFs and scanned images, ensuring robust processing of various content types.

At its core, the solution employs a llama model, meticulously trained on specialized PII datasets, to detect and classify sensitive information into categories such as PII data. This approach ensures precise and context-aware redaction with minimal error rates. Sensitive data is effectively masked using black rectangles, preserving document usability and structure. The system delivers redacted outputs in user-specified formats, such as PDFs or text files, catering to diverse operational requirements. By reducing manual effort, it minimizes the risk of data breaches, enhances operational efficiency, and ensures compliance with critical privacy regulations, including GDPR and HIPAA.

This project represents the seamless integration of OCR and machine learning technologies, offering a scalable, adaptable, and user-friendly solution for industries such as healthcare, finance, and legal services. It sets a new standard for data privacy and document security in today's increasingly digital world.

**CONTENTS**

| **S. NO.** | **TITLE** | | **PAGE NO.** |
| --- | --- | --- | --- |
|  | **ABSTRACT**  **TABLE OF CONTENTS**  **LIST OF FIGURES**  **LIST OF TABLES** | | i  ii  iii  iv |
| **1** | **Introduction** | | 1-3 |
| **2** | **Literature Survey** | | 4-7 |
|  | 2.1 | Research Gap | 8 |
| **3** | **Proposed Work Architecture, Technology Stack, Implementation Details** | | 9-15 |
| **4** | **Results & Discussions** | | 16-21 |
| **5** | **Conclusion & Future Scope** | | 22 |
| **6** | **References** | | 23 |

**LIST OF FIGURES**

| **Fig. No.** | **Figure Name** | **Page No.** |
| --- | --- | --- |
|  |  |  |
| 1 | Architecture | 13 |
| 2-8 | Output | 18-21 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**LIST OF TABLES**

| **Table No.** | **Table Name** | **Page No.** |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

1. **INTRODUCTION**

Redaction is the act of hiding or deleting sensitive, private or confidential information from documents, data sets, images, and videos, or other sources before they are disclosed publicly. The aim of redaction is to protect sensitive data, such as names , addresses, phone numbers, financial information, trade secrets, or classified information and materials, from unwanted usage or access. This practice is widely applied in a number of industries such as law, healthcare, government, and business. In general, redaction is done by identifying sensitive elements by screening the data so that the remaining content maintains its usability and integrity. Modern redaction tools increasingly rely on rule-based techniques, predefined patterns and algorithms to detect and redact sensitive information.

Redaction protects sensitive information through systematic obscuring or erasure of data that might result in privacy violations, security issues, or competitive disadvantages. From removing personal identifiers from public records to obscuring classified information in governmental documents, from anonymizing proprietary information in business reports to anything else, redaction ensures that non-sensitive content alone remains accessible. This would allow data usability while also ensuring it follows all applicable privacy regulations and industry guidelines. In today's world where there is increased digital communication and data sharing, there has never been a greater time to be redaction effective.

The failure to implement appropriate redaction measures can result in a chain of serious consequences across a range of domains. Un-redacted information will expose organizations and individuals involved to risks such as hacking of data, identity thefts, financial frauds, among other things, besides unlawful divulgence of classified information. In legal settings, such cases may lead to the breaching of attorney-client privilege while infringing on data protection like in the case of General Data Protection Regulation. In corporate environments, exposure of un-redacted proprietary information could lead to intellectual property theft or give competitors an unfair advantage. At its most critical, leakage of sensitive governmental or defense-related information could compromise national security or individual lives. Therefore, redaction is a basic tool for reducing these risks, maintaining compliance with data protection regulations, and fostering trust in an increasingly interconnected world.

Redaction tools often use rule-based approaches as a primary method for detecting and redacting sensitive information in different formats. Such techniques rely on predetermined rules, patterns, and logic to identify specific kinds of data that need redaction. For instance, they can use regular expressions, or regex, to determine structured information such as credit card numbers, Social Security numbers, or email addresses due to their known patterns. Another function that the keyword-based rules offer is to flag sensitive terms or phrases, names of persons, classified titles of project, or even legal terminology. Once they have detected these, then there are a variety of techniques implemented by the tools: masking or replacing sensitive contents with some placeholders so the redacted information cannot be read or otherwise exposed.

Structured and semi-structured data prove easy in rule-based redaction methods since patterns here are usually known and consistent. Such systems work well in environments with controlled information types-such as financial records or legal documents or government forms-whose type of data are well defined. Organizations can tailor the rules that make up these systems very flexibly, either to include or exclude data of particular types. They have, however, significant effort when trying to configure and maintain them. For instance, creating comprehensive and accurate rules demands a deep understanding of the data and the potential variations of sensitive elements within it.

While rule-based tools are efficient and deterministic, they can face challenges when dealing with unstructured or complex data, such as free-text fields or documents with subtle contextual nuances. In such cases, the rigidity of rule-based systems may result in either over-redaction, removing information that is not necessary, or under-redaction, missing critical sensitive data. Rule-based redaction, despite its limitations, remains a strong and reliable approach where precision, predictability, and control are valued. Rule-based techniques, therefore, form a critical part of protecting privacy and maintaining compliance with data protection regulations because they ensure systematic identification and cover-up of sensitive data.

The project utilizes a state-of-the-art pre-trained language model, **Meta-Llama-3.1**, to efficiently identify and redact sensitive information in text documents. **Meta-Llama-3.1** is a highly capable large language model designed for a wide range of natural language processing tasks, offering robust performance due to its extensive pre-training on diverse datasets.

In this project, this models are fine-tuned using **LoRA (Low-Rank Adaptation)** to specialize them in detecting and classifying sensitive information such as PII, financial data, and other confidential elements. By leveraging **4-bit quantization** through the **Unsloth framework**, the model achieves significant memory savings while maintaining high accuracy. This allows the system to handle complex tasks, such as processing scanned documents or forms, and securely redact sensitive details before returning structured outputs in user-specified formats. This model forms the backbone of a scalable and efficient automated redaction pipeline tailored to meet privacy and compliance standards.

1. **LITERATURE SURVEY**

Redaction involves a range of methods and techniques aimed at identifying and concealing sensitive information in documents, images, videos, or datasets. These techniques range from manual approaches, where individuals review and redact content, to advanced automated systems. Initially, redaction relied on simple manual processes, such as blacking out text with markers or removing sections of documents. While effective in small-scale or straightforward scenarios, manual redaction is time-consuming, error-prone, and unsuitable for handling large volumes of data.

The rise of digital tools has enabled the development of automated redaction techniques that are more accurate and efficient. Among these, rule-based techniques are considered the traditional approach. Rule-based systems use predefined patterns and logic to detect sensitive information. For example, regex can identify structured data like Social Security numbers, credit card information, dates, or phone numbers based on their predictable formats. Similarly, keyword-based rules can scan for certain terms, phrases, or identifiers associated with confidential information. Once detected, the data is masked, replaced, or removed to ensure that it cannot be accessed.

More advanced redaction tools employ techniques such as context-aware algorithms and machine learning, which allow analysis of surrounding data to comprehend context and improve the detection of sensitive content. Yet, rule-based techniques form a foundation of automated redaction due to their simplicity, reliability, and ease of implementation. They are very efficient in the case of structured and semi-structured data and are used widely in industries such as law, finance, and government.

Despite these benefits, rule-based methods are not particularly effective in handling unstructured or context-dependent data where sensitive information does not necessarily follow any predictable patterns. In this case, hybrid approaches, which combine rule-based systems with artificial intelligence, will be more complete. Nevertheless, rule-based redaction remains a vital and practical technique to ensure data privacy and compliance with regulations across a broad range of applications.

Peng et al. [1] explores the integration of artificial intelligence into the redaction process to enhance the accuracy and efficiency of protecting sensitive information. Using controlled experiments, the researchers compared three methods: manual redaction, classical machine learning (ML)-based tools, and the AI-driven iDox.ai Redact tool, to compare their performance in terms of accuracy and time. Manual redaction, done through Adobe Acrobat, was strictly human effort-based, while the classical ML tool relied on pre-programmed algorithms to detect patterns but required manual input for data that were not recognized. In contrast, iDox.ai Redact used advanced AI algorithms to automate the identification of sensitive data, adapting to diverse patterns with minimal human intervention. The results showed that iDox.ai Redact significantly outperformed the other methods, achieving the highest accuracy (97.10%) and the fastest processing time (15.75 minutes per document), demonstrating its effectiveness in reducing errors and handling complex datasets.

However, the study also highlights several challenges and limitations. AI tools, including iDox.ai Redact, face difficulties in recognizing diverse or highly complex sensitive data types, especially when patterns deviate from standard formats or require contextual understanding. These tools often require customization to meet specific industry or regulatory requirements, limiting their out-of-the-box adaptability. Besides this, the AI tools depend greatly on the quality and variety of the training data for producing high accuracy while dealing with rare or specialized data sets. Additionally, a very big area that is left for development includes the user experience factor where seamless workflows and clear-cut interfaces will ensure further deployment in multiple applications. Improvement on these challenges is likely to make full realization possible on the applications of AI-driven redaction tools.

Hill et al. [2] the paper "On the (In)effectiveness of Mosaicing and Blurring as Tools for Document Redaction" explores the vulnerabilities of using mosaicing (pixelation) and blurring as methods for obscuring sensitive text. Using Hidden Markov Models (HMMs), the study shows how these redaction techniques leave behind residual information that can be exploited to recover the original text, using methodologies inspired by automatic speech recognition. The researchers applied the said redaction methods to images of text with varying typeface, font size, size of mosaic grid, and blur radius. Pixel patterns of the redacted text were then observed for analysis by applying the techniques of sliding windows and vector quantization like k-means clustering. HMMs were subsequently learned on rendered text for obtaining the hidden sequences and ultimately for recovering the original data. The results indicated that HMMs are very effective for text reconstruction, even with noise or degradation conditions, and strongly outperform previous brute-force heuristics, especially challenging redaction scenarios. More real-world examples were then used to demonstrate the failure of mosaicing and blurring as forms of secure redaction. The study concludes the techniques to be unreliable due to the residual data they tend to leave behind, which recommends the adoption of stronger methods of redaction, thereby replacing sensitive text with a placeholder or solid black bars so that it is secure and not likely to be recovered. The paper highlights several challenges and drawbacks in using these techniques for secure text redaction. One major challenge is the residual information left behind after mosaicing or blurring, which can be exploited to recover the original text. These techniques do not sufficiently obscure the underlying structure or patterns of the text, allowing tools like Hidden Markov Models (HMMs) to model and reconstruct the redacted content effectively.

Another limitation is that each technique relies on specific parameters for redaction, including grid size in mosaicing or blur radius in blurring. If such parameters are not aggressive enough, it is likely to significantly increase the risk of recovery, but overly aggressive settings degrade the usability and readability of the document. These methods are also intrinsically susceptible to sophisticated reconstruction algorithms that can learn to adapt to varying redaction conditions, such as noise or degradation, as shown in the paper.

The results also show that these techniques fail miserably when applied to text that has uniform patterns or predictable structures, such as printed or formatted documents, for which recovery tools work fantastically well. The last concern is that these methods are unreliable and not robust enough for use in the real world, raising questions about whether they should be applied in situations where data privacy and security are paramount. These concerns highlight the requirement for stronger redaction methods, including solid masking or content replacement, to ensure that sensitive information cannot be recovered.

Bland and Maxwell[3]  reveals important vulnerabilities in processes of redaction of a PDF by demonstrating how information about glyph positioning, sub pixel offsets, for instance, can leak redacted text. The research establishes that slight spatial differences of glyph widths, positions, and shifts in a document can be used to reverse-engineer hidden text, especially with structured or predictable datasets such as names, numbers, or addresses. Using statistical and heuristic methods, the researchers analyzed positional patterns retained in the metadata of redacted documents. They evaluated 11 widely used PDF redaction tools and discovered that most retained glyph metadata even after the visible text was obscured. This allowed for the recovery of sensitive information, proving the insufficiency of current methods. To demonstrate impact, the team applied their attack to publicly available datasets, for example, voter lists; they successfully reconstructed sensitive details from subpixel differences. Responding to the researchers were countermeasures, namely advanced redaction techniques fully eliminating positional metadata and even suggested changes in document handling workflows. The team also involved tool developers to address these risks as well as enhance the security of PDF redaction tools.

Despite its great contributions, the study highlights several challenges and drawbacks in the current redaction landscape. One of the major issues is that metadata in PDF files is persistent and remains intact even when visible content is obscured, creating a hidden vulnerability that is difficult to address without modifying fundamental file structures. The use of heuristic methods and glyph positions further shows how little design oversights in tools for redaction can be hijacked and used to reveal sensitive information on a scale. This study also gives an indication of the great challenge that is educating and informing users and organizations because most of them assume such obfuscation is enough. The main disadvantages include the complexity of effective countermeasures, which are mostly redesigns of existing tools and workflows, often at the cost of usability or performance. These issues emphasize the urgent need for more powerful and comprehensive redaction practices that address not only visible content but also underlying metadata to ensure complete data security.

Osipenko[4] presents an overview of the prevalence, types, and impacts of data redaction in National Institute for Health and Care Excellence (NICE) technology appraisals (TAs) and highly specialized technology evaluations (HSTs) over two decades. It was found that 82% of the analyzed appraisals contained redacted data, including clinical, economic, and pricing information. Of importance, all HSTs and 91% of single TAs (STAs) contained redacted data, as opposed to 59% of multiple TAs (MTAs). Clinical data were redacted in 65.7% of TAs and all HSTs, and pricing data were obscured in 58.3% of TAs. Moreover, AEs and QALYs were redacted in 31.4% and 34.6% of TAs, respectively. The study shows the trend of redactions to grow over time, with each appraisal since 2019 having data redacted. From the data, it was obvious that redactions were more commonly found in oncology-specific appraisals (87% of those) than elsewhere (78% of nononcology appraisals).

These practices had significant impact, with confidentiality claims—a commercial or academic one generally—made without clear cause for restriction, severely impairing transparency. The inability to access the information easily did not allow clinicians, patients, or researchers to make proper judgments or to get necessary critical clinical information. It was also evident from the study that NICE appraisals censored the clinical trial details and adverse event data but they were public through other regulatory agencies, including Health Canada or the European Medicines Agency. Policy changes for transparency have been urged through the paper. Recommendations for reform would be to clarify guidelines on redaction, un redaction of academic-in-confidence data post-publication, and promote international cooperation for harmonization in practices related to data transparency. In conclusion, the lack of transparency in NICE documentation represents a substantial failure on the part of this organization regarding the commitment it has shown toward evidence-based decision-making and patient-centered care, calling for urgent reform to achieve accountability and accessibility in health technology assessments.

**2.1 RESEARCH GAP**

Several critical research gaps in redaction technologies are directly applicable to enhancing our automated redaction system. A prominent gap is the adaptability of AI models to diverse and complex datasets. Many existing solutions depend heavily on high-quality, domain-specific training datasets, which limits their effectiveness in handling unstructured or rare sensitive data. For our project, developing AI models capable of generalizing across varied contexts without extensive retraining would significantly enhance our system's scalability and flexibility, ensuring it performs reliably in diverse real-world scenarios.

Another significant challenge is the elimination of residual vulnerabilities left behind in redacted documents. Techniques such as mosaicing, blurring, or glyph metadata often leave traces of information that sophisticated algorithms can exploit to reconstruct sensitive content. For example, metadata in PDFs, such as glyph positioning or subpixel offsets, poses a risk of data leakage even when visible content appears obscured. Addressing this issue by integrating robust methods to securely remove residual data and underlying structural weaknesses will ensure comprehensive and end-to-end redaction security in our system.

Improving the usability and user experience of redaction tools is another critical area. Many current systems feature complex workflows or unintuitive interfaces, making them difficult to use, especially for non-technical users. For broader adoption, it is essential that our redaction system incorporates simplified workflows, clear guidelines, and user-friendly designs to make redaction more accessible and efficient.

Furthermore, the lack of robust solutions for redaction in dynamic formats such as videos, audio, or real-time datasets represents an opportunity for innovation. As data formats become increasingly varied, expanding our system to support these evolving data types will position it as a cutting-edge, comprehensive tool for sensitive information management.

Lastly, balancing transparency and confidentiality in data redaction remains a significant challenge. For example, studies on NICE appraisals reveal that excessive redaction without clear guidelines can hinder accessibility and trust. Developing policy frameworks that harmonize data protection with transparency, while ensuring compliance with regulations like GDPR and HIPAA, would strengthen the reliability and usability of redaction systems.

Addressing these gaps will enhance the robustness, security, and user-friendliness of our system, ensuring it is well-equipped to meet the challenges of sensitive data management across industries such as healthcare, finance, and legal services.

1. **PROPOSED WORK**

Our project entails the development of a Flask-based web application that presents a robust and automated solution for redacting sensitive information in PDF documents. In a world where data protection, such as PII and other classified information, has become a key concern, the application addresses a growing need in healthcare, finance, legal, and government institutions for secure and efficient data redaction. The cutting-edge natural language processing models, including the Llama model, are integrated with PDF processing libraries like PyMuPDF, Tesseract OCR and pdfplumber to provide a scalable, user-friendly solution.

In today’s data-driven world, protecting sensitive information in documents has become an indispensable requirement for industries such as healthcare, finance, legal, and government institutions. These sectors handle a vast amount of Personally Identifiable Information (PII) and classified details that demand secure handling to prevent data breaches and ensure compliance with stringent regulations like GDPR and HIPAA. To address this critical need, our project delivers a state-of-the-art Flask-based web application that automates the redaction of sensitive information from documents such as PDFs and scanned images. By combining advanced natural language processing (NLP), optical character recognition (OCR), and document-processing technologies, the system provides an efficient, user-friendly, and highly scalable solution to a traditionally labor-intensive task.

The application begins by offering users a seamless document upload interface. Users can submit files in diverse formats, including PDFs and scanned images, ensuring compatibility with a variety of real-world use cases. Uploaded files are temporarily stored for processing, adhering to best practices for efficient resource utilization and data privacy. The initial stage involves extracting textual content and layout information from these documents. This is achieved using libraries like PyMuPDF and pdfplumber, which work in tandem to capture both text and its spatial metadata, including bounding boxes. These spatial coordinates are critical for mapping each word’s exact location on the document, enabling precise identification and targeting of sensitive content. This capability is particularly essential for handling complex layouts such as multi-column documents, tables, and forms, where traditional methods often struggle.

To identify sensitive information, the application leverages a pre-trained language model Meta-Llama-3.1, which is fine-tuned using Low-Rank Adaptation (LoRA) techniques. This model has undergone extensive customization to meet the specific requirements of sensitive data detection, with their efficiency further enhanced through 4-bit quantization using the Unsloth framework. This approach allows the models to process unstructured text efficiently, even in resource-constrained environments, while maintaining exceptional accuracy. The models analyze the extracted text, classifying tokens into predefined categories such as names, addresses, phone numbers, and financial data. Through their advanced contextual understanding, they can discern subtle differences in meaning, such as distinguishing between “John” as a name and “John Street” as an address. Additionally, these models excel in grouping multi-token entities into coherent strings, ensuring that entities spanning multiple lines or locations within a document are accurately recognized and classified.

The sensitive information detected by the models is presented to users in a structured format, offering transparency and allowing for easy review and customization. Users can refine the redaction process by adding or excluding specific content based on their needs, striking a balance between automation and user control. Once the sensitive entities are identified and reviewed, the application proceeds to the redaction phase. This involves securely obscuring the identified content by superimposing black rectangles over the sensitive areas. These redactions are applied directly to the document, ensuring that the sensitive information is irreversibly removed and cannot be retrieved or reverse-engineered. The system’s ability to handle diverse layouts and complex structures makes it robust and versatile, catering to a wide range of document types and scenarios.

The application’s functionality is facilitated through two primary endpoints: the /api/PDFpreprocess endpoint, which preprocesses the document and provides a detailed JSON response containing extracted text, bounding boxes, and detected entities, and the /redact\_pdf endpoint, which performs the actual redaction based on user inputs and the model’s detection results. This architecture not only ensures precision and efficiency but also enables seamless integration into existing workflows, enhancing the system’s scalability and adaptability.

Designed with scalability in mind, the application is capable of processing large volumes of data without compromising performance. Temporary file handling optimizes resource utilization, while the integration of libraries like PyMuPDF, pdfplumber, and Tesseract OCR(Optical Character Recognition) ensures precise text extraction and manipulation. The use of advanced NLP models further enhances the system’s ability to process unstructured text, making it robust even in challenging scenarios. This scalability and efficiency make the application suitable for real-world applications in sectors where data security and compliance are critical.

In healthcare, for example, the system can redact patient information to ensure compliance with HIPAA regulations, protecting sensitive data while maintaining the usability of medical records. In the financial sector, it can secure confidential financial documents, mitigating the risk of data breaches and safeguarding sensitive information. Similarly, in legal settings, the application can process case files while protecting client confidentiality, streamlining operations without compromising data security.

To further improve the system, future enhancements will focus on expanding the entity recognition capabilities of the language models to accommodate a broader range of sensitive information categories. Optimization of the application’s processing speed will ensure faster handling of large-scale operations, while the development of a more dynamic and responsive user interface will enhance the user experience. These improvements will solidify the application’s position as a comprehensive and versatile solution for automated document redaction, capable of meeting the evolving needs of diverse industries.

This project represents a powerful integration of advanced technologies, combining the capabilities of NLP, OCR, and document processing into a cohesive and user-friendly system. By automating the redaction of sensitive information, the application significantly reduces manual effort, minimizes errors, and enhances data security. It addresses the growing challenges of data privacy in an increasingly digital world, offering a reliable, efficient, and scalable solution for industries that handle confidential information.

**3.1 Data Collection & Preprocessing**

The first step of our proposed solution involves the collection of data, which we approached by utilizing a secondary dataset specifically curated for our project. For this purpose, we chose the **PII Masking 300k** dataset from Hugging Face, a resource specifically designed to facilitate the detection and masking of personally identifiable information (PII). This dataset provides an extensive collection of text samples annotated with detailed metadata to support tasks like entity recognition and redaction. The primary columns in the dataset include the text column, which holds raw textual data, and the entities column, which lists sensitive entities such as names, email addresses, phone numbers, and other PII types. Additionally, the dataset contains start and end indices for each entity, which pinpoint the exact locations of these sensitive elements within the text. These annotations make it highly suitable for training machine learning models aimed at detecting and redacting PII with precision. The dataset's core objective aligns perfectly with our research goals, as it is designed to aid in developing robust models for privacy preservation and sensitive data handling.

Our team conducted an in-depth examination of the dataset and found it to be a valuable foundation for building an automated redaction system. However, like any real-world dataset, it was not without its challenges. Upon closer inspection, we identified several impurities and inconsistencies that required preprocessing. Some text samples were missing annotations or contained partial labeling, which could introduce noise into the training process. Additionally, there were instances of duplicate entries and text samples unrelated to PII, which needed to be filtered out. Formatting inconsistencies, such as irregular text encodings and the presence of special or non-standard characters, were also observed. These issues posed a potential risk to the reliability and accuracy of the models we aimed to train, necessitating a meticulous data cleaning and preprocessing pipeline.

To address these challenges, our team implemented a series of preprocessing steps designed to ensure the dataset's quality and suitability for model training. We began by standardizing the text format to ensure consistency across all entries, which involved normalizing character encodings and removing any extraneous or invalid characters. Duplicate entries were identified and removed to prevent redundancy, while irrelevant text samples not associated with PII were filtered out to reduce noise. A key aspect of preprocessing involved validating the alignment of start and end indices with their corresponding substrings in the text column. Custom validation scripts were written to cross-check these indices and rectify any discrepancies. Furthermore, columns that were not directly relevant to the task of PII detection, such as additional metadata fields, were eliminated to streamline the dataset and focus solely on the necessary features.

In addition to these steps, we conducted exploratory analysis to better understand the dataset's structure and features. This included visualizing word frequency distributions to identify common patterns and assess the diversity of sensitive entities present in the dataset. Such analyses helped us gain insights into the dataset's composition, which informed our feature engineering and model design. By addressing the impurities and preparing the data in a structured format, we ensured that it provided a robust foundation for training models capable of accurately detecting and redacting PII. The comprehensive preprocessing pipeline not only improved the quality of the data but also enhanced its usability for machine learning tasks. With the cleaned and refined dataset, our system is well-positioned to deliver a reliable and efficient automated redaction solution that can handle a wide range of real-world documents.

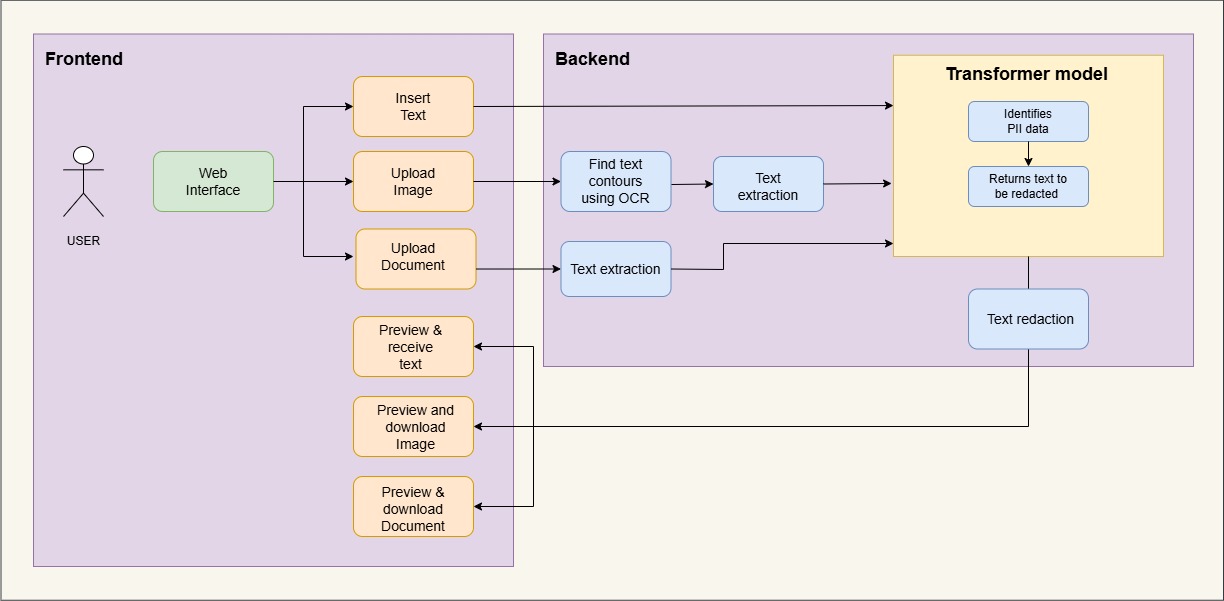


Fig. 1**: ARCHITECTURE**

**Above Fig. 3.1**: The automated redaction system architecture and workflow demonstrate three primary components of the system: the frontend user interface, backend processing, and machine learning model integration. These components work cohesively to ensure high accuracy and efficiency in the redaction process.

The **frontend interface** is a user-friendly, interactive, and web-based platform through which users can upload various types of files, including PDFs, scanned images, and text documents. This interface not only facilitates the uploading process but also provides real-time previews of the uploaded content and its redacted version, enabling users to visualize the final output before downloading it. The frontend design ensures smooth navigation and a step-by-step guide, leading users from document upload to downloading the finalized redacted document, available in PDF or plain text formats.

The **backend processing** serves as the system's core, handling document analysis and redaction. Once a document is uploaded, advanced OCR technology, powered by Tesseract.js, is employed to extract text content and retrieve word bounding box coordinates. This allows the system to process complex document layouts such as multi-column text, tables, forms, and even handwritten content with high precision. The extracted text and spatial metadata are then passed through a pre-trained language model, such as Meta-Llama-3.1, fine-tuned for sensitive entity detection. These models, customized using LoRA techniques and optimized for efficiency through 4-bit quantization, analyze the text to identify and classify sensitive information.

The models tokenize the text, detect entities such as names, phone numbers, addresses, and financial details, and classify them into predefined categories of sensitive information. Leveraging advanced contextual understanding, the models ensure accurate detection while minimizing false positives and negatives. They also identify multi-token entities spread across lines or sections, ensuring a comprehensive understanding of sensitive data in the document.

Once the sensitive information is detected, the system transitions to the **redaction phase**. The backend securely masks the identified data by overlaying black rectangles ensuring that the redacted content cannot be selected, copied, or retrieved. Users can also customize the redaction process by manually searching for and specifying additional words or phrases to redact, providing flexibility and ensuring that no sensitive data is missed.

The final redacted document is presented to users for inspection and download in their preferred format. The backend processes ensure a seamless and efficient workflow, combining Tesseract.js, PyMuPDF, pdfplumber, and fine-tuned language models to deliver a robust and scalable redaction solution. The system's adaptability makes it ideal for diverse applications across industries such as healthcare, finance, legal, and government sectors.

1. **RESULTS**

The main purpose of this project is to develop a highly reliable and automatic system for redacting sensitive information in documents, offering a user-friendly interface flexible enough to meet the diverse needs of users. This system is designed to detect and redact Personally Identifiable Information (PII) and other confidential data from PDFs, scanned images, and text files. By utilizing cutting-edge technologies such as Tesseract OCR for text extraction and the Llama model for entity recognition, the application automates a traditionally manual and error-prone task. The platform balances automation and user-driven customization to ensure efficiency, adaptability, and compliance with strict data protection regulations such as GDPR, HIPAA, and CCPA. The result is a fully functioning web application that provides an efficient, intuitive, and scalable solution for automated redaction. It is tailored for use in industries such as healthcare, finance, legal, and government sectors. The system supports a wide range of document formats including PDFs, scanned images, and plain text files, making it applicable across different use cases.

The system begins with an interactive user interface that allows users to easily upload documents. After uploading, the document is processed using Tesseract OCR to extract text content, word frequencies, and spatial coordinates, accurately mapping where each word falls within the document. This capability ensures that complex layouts such as tables, forms, and multi-column formats can be handled effectively. The extracted data, including text and metadata, is then passed to the Llama model, a powerful transformer-based machine learning model pre-trained to detect and classify sensitive information.

The Llama model analyzes the extracted text while considering its contextual relevance. It identifies sensitive entities, such as names, phone numbers, addresses, financial details, and health-related data. By leveraging deep learning capabilities, the model minimizes false positives and negatives, ensuring high accuracy in detection. Once sensitive entities are identified, the system presents a real-time preview of the document on the upload page, enabling users to see the redacted output before finalizing the changes. The real-time preview feature enhances transparency, as users can verify the redaction results and adjust them if necessary. Additionally, users can search for specific words or phrases that they may want to redact manually, ensuring no sensitive information is overlooked.

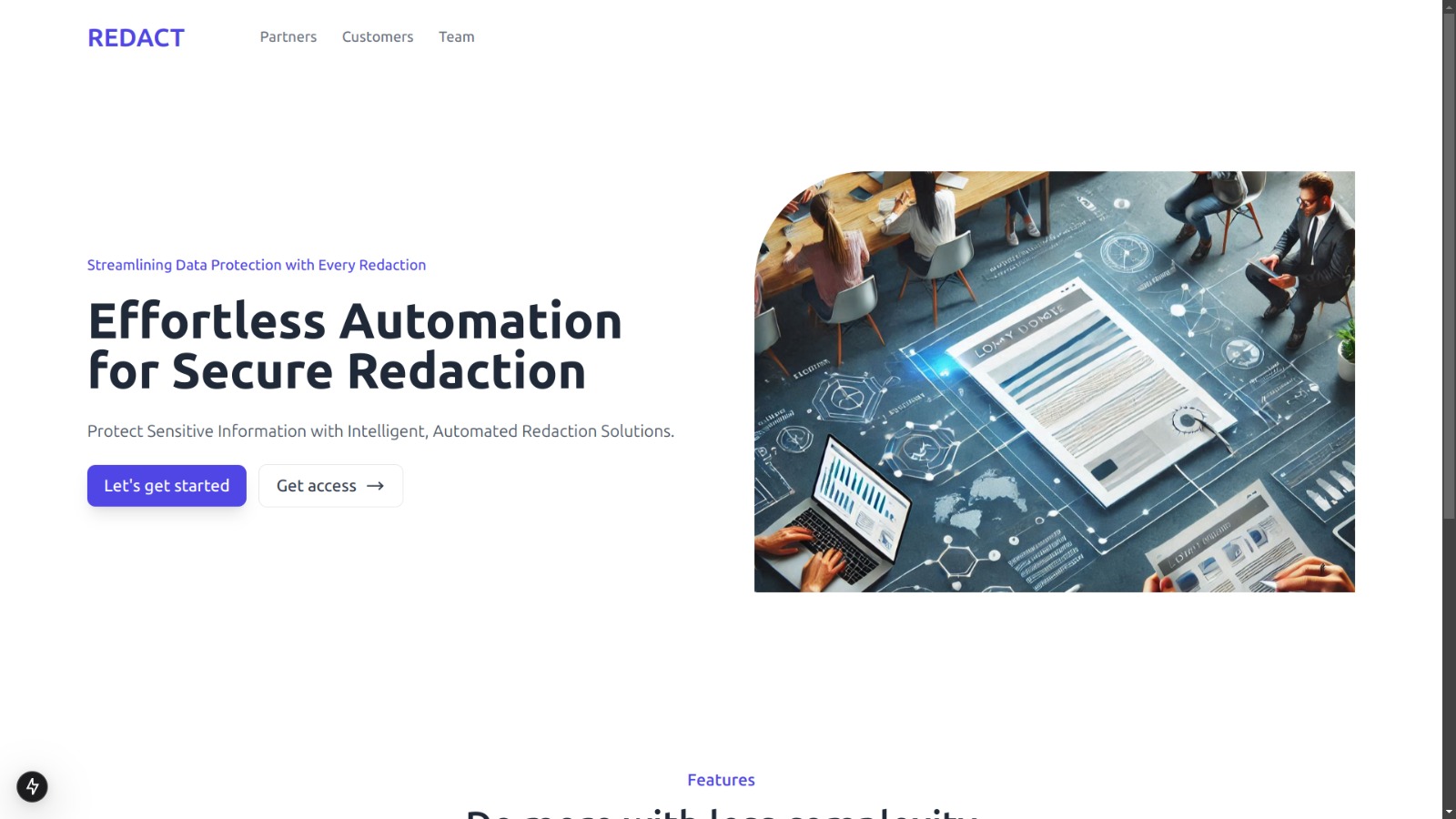
During the redaction phase, the system uses secure masking techniques, such as overlaying black rectangles or replacing text with placeholders, to ensure the redacted content cannot be recovered or accessed. After reviewing the document, users can finalize the redaction process, and the system generates the redacted document in their desired format (PDF, Word, or plain text). This process preserves the structure and integrity of the original document while ensuring that sensitive information is securely hidden.

Several advanced libraries and frameworks were utilized during the development of the system. The frontend was built using Next.js, providing a dynamic and responsive user experience. The backend, developed in Node.js, manages server-side operations and handles data processing. Text extraction was carried out with Tesseract.js, a JavaScript library for OCR, which ensures seamless processing of uploaded files. Sensitive entity detection is performed by the Llama model through an API endpoint, with Express.js managing the routes and data flow through the API.

The development of the Llama model required significant effort to fine-tune it for the detection of sensitive information in various document types. Extensive datasets were used to train the model, and several optimization steps were taken to ensure that the system could handle large-scale processing without compromising performance or precision. Despite the challenges, the system was refined and optimized to provide a robust and reliable solution.

In summary, this project combines advanced OCR technology, transformer-based NLP (via the Llama model), and modern web development tools to create a scalable, secure, and versatile redaction platform. The system prioritizes user experience and customization, representing a significant advancement in safeguarding sensitive information. This solution meets the growing demand for automated, reliable data protection across various sectors, including healthcare, finance, legal, and government.

Our work stands out by leveraging the Llama model, a sophisticated transformer-based NLP tool, for detecting and classifying sensitive information in text. Unlike conventional redaction systems that rely on manual processes or static, rule-based approaches, our system provides a fully automated solution with customizable features. This automation ensures high accuracy and efficiency while handling complex document layouts. Additionally, the system supports multiple document formats and allows for real-time preview and manual search, offering users a unique blend of automation and customization. The ability to work with PDFs, scanned images, and text files in various formats makes the system adaptable for a range of industries. By delivering a comprehensive, user-friendly solution, our project sets a new standard in automated document redaction.

Fig.2:  **Web application interface**

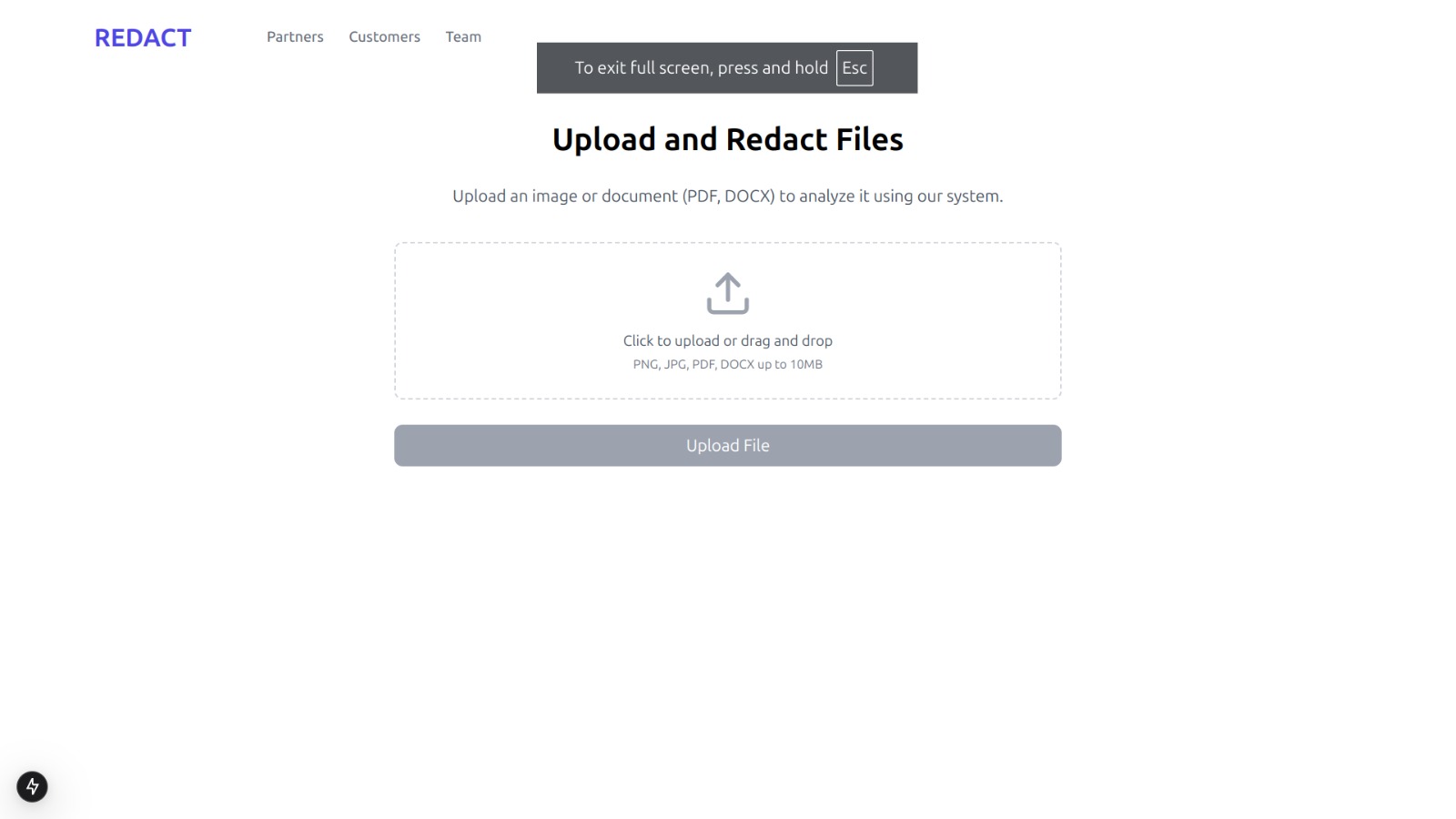
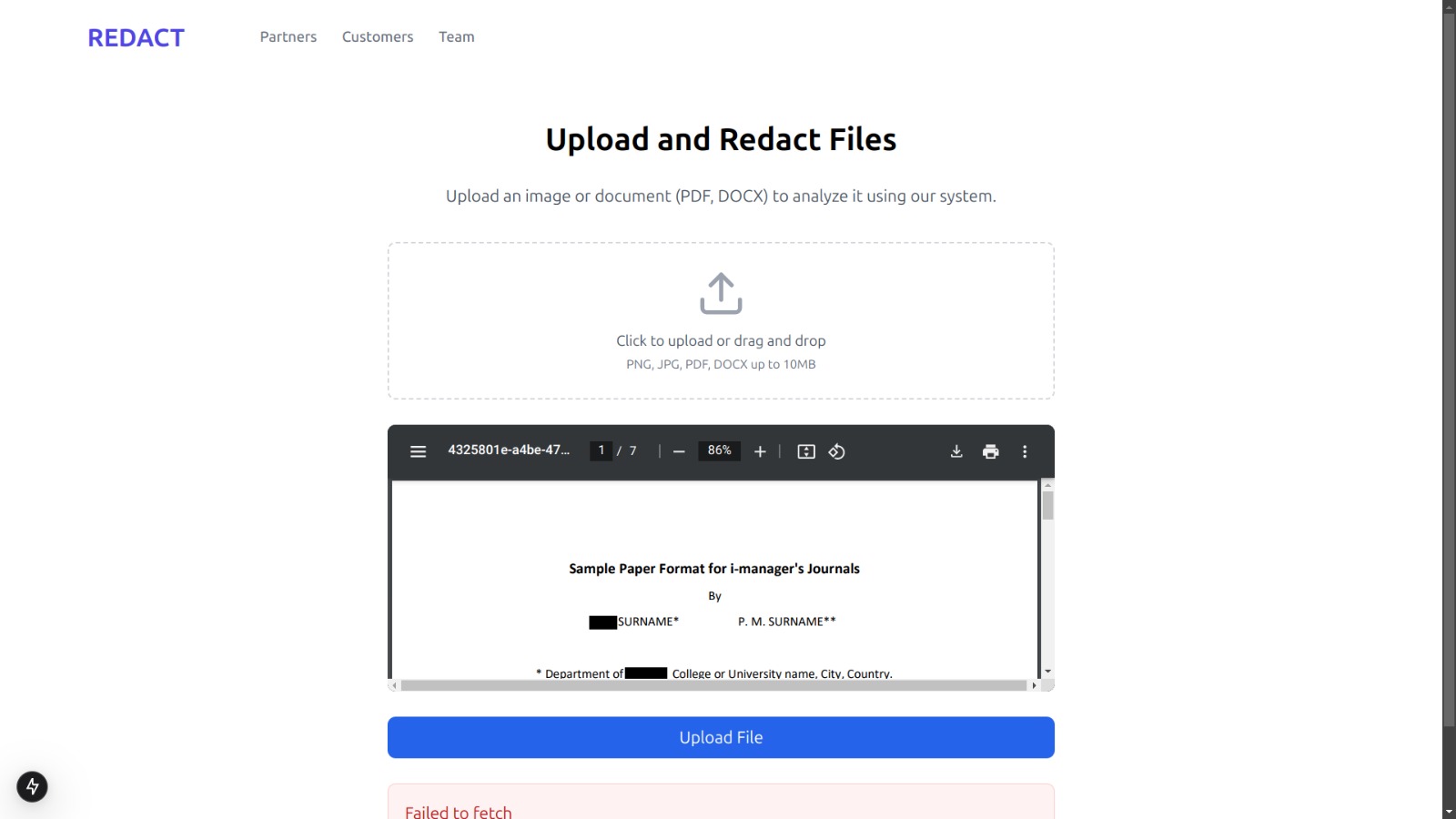
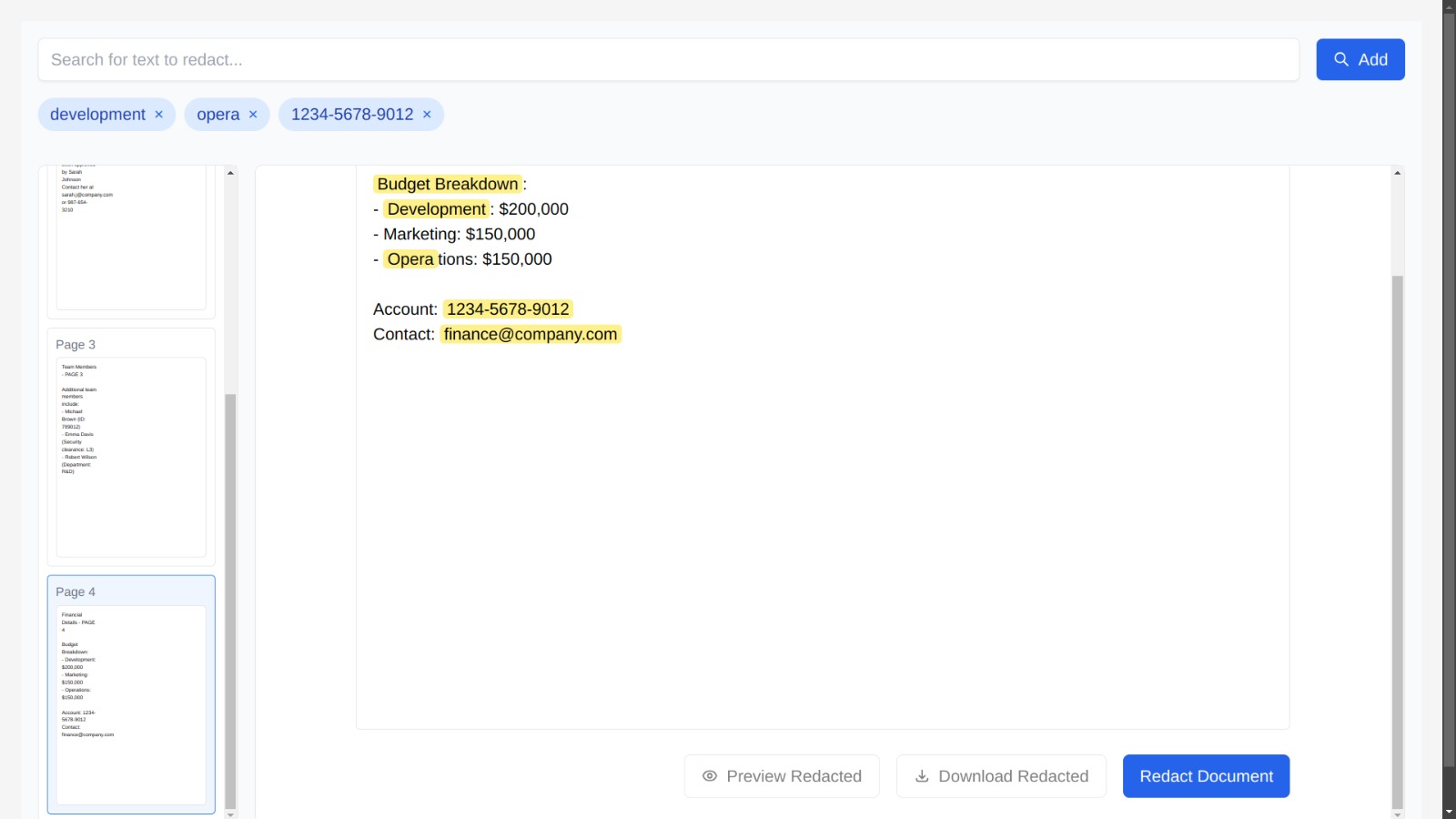
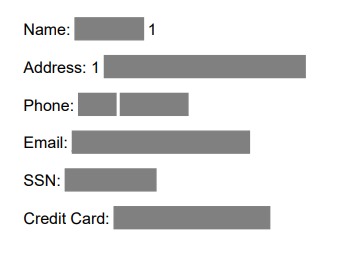
****

Fig. 3:

****

**F**ig:4

****

**** ****

****

**Fig 8:** Final output

1. **CONCLUSION**

The developed project successfully achieves its goal of creating an application that leverages the web platform to automatically redact documents, ensuring that sensitive information remains securely masked without compromising the usability or integrity of the document. By utilizing Tesseract OCR for precise text extraction and the Llama transformer model for accurate sensitive data identification, the system provides a user-friendly yet highly efficient solution for redacting confidential information.

A key challenge encountered during the project was the training and fine-tuning of the Llama model. This process was crucial for achieving high accuracy in sensitive entity detection and classification. It involved extensive testing, iterative optimization, and careful dataset refinement. These efforts ultimately improved the system’s performance and reliability, enabling it to effectively handle a wide range of sensitive data types.

The project holds significant promise for real-world applications in sectors such as healthcare, finance, legal, and government, where data confidentiality is paramount. The system’s ability to manage multiple document formats and structures, along with advanced redaction capabilities and customization options, makes it adaptable to diverse user needs. While the current system is robust and fully functional, there are opportunities for future improvements. These include expanding the categories for entity recognition, optimizing processing speeds, and enhancing the user interface to increase interactivity and overall user experience.

In conclusion, this project not only addresses the growing demand for automated data privacy solutions but also establishes a strong foundation for future advancements in document security and redaction technologies. It exemplifies the successful integration of state-of-the-art OCR and machine learning technologies into a scalable, adaptable solution for protecting sensitive information, setting the stage for continued innovation in this critical area.

**REFERENCES**

1. Peng, Sida, Ming-Jen Huang, Matt Wu, and Jeremy Wei. "Transforming Redaction: How AI is Revolutionizing Data Protection." *arXiv preprint arXiv:2409.15308* (2024).
2. Hill, Steven, Zhimin Zhou, Lawrence Saul, and Hovav Shacham. "On the (in) effectiveness of mosaicing and blurring as tools for document redaction." *Proceedings on Privacy Enhancing Technologies* (2016).
3. Bland, Maxwell, Anushya Iyer, and Kirill Levchenko. "Story beyond the eye: glyph positions break PDF text redaction." *arXiv preprint arXiv:2206.02285* (2022).
4. Osipenko, Leeza. "Audit of data redaction practices in NICE technology appraisals from 1999 to 2019." *BMJ open* 11, no. 10 (2021): e051812.