LLM-Assisted Detecting and Redacting Confidential Information for Government Information Disclosure

Masaki Hasegawa

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

Shin'ichiro Matsuo, Chair
Wenjing Lou, Co-chair
Melissa Cameron

April 25, 2025
Arlington, Virginia

Keywords: Large language model, Public Sector, Government Process Optimization.

LLM-Assisted Detecting and Redacting Confidential Information for Government Information Disclosure

Masaki Hasegawa

(ABSTRACT)

Generative AI, especially large language models (LLMs), has advanced rapidly, with real-world applications growing steadily. However, the use of generative AI in the public sector has lagged behind the private sector. This paper focuses on the "Governmental Information Disclosure Process," which is vital in democratic countries' administrative systems. Many developed nations require government agencies to disclose information to citizens, excluding confidential data such as personal information. Although agencies must confirm the presence of confidential information and redact or mask it before release, this process is still manual, creating significant room for improvement. Additionally, since the information to be masked is defined in natural language, such as legal text, interpreting documents' contexts to determine what qualifies as confidential is resource-intensive. In this context, LLMs, capable of inferring context and general knowledge, could efficiently identify parts of documents that require masking. This paper first reviews the existing literature on sensitive or confidential information detection using LLMs, clarifying the use cases and the category of information identified in both the private and public sectors. Then, as a case study, we create sample documents modeled after Japanese administrative texts and compare the detecting and masking results performed by testers with administrative experience, following legal requirements, with those generated by an LLM. This study contributes by proposing end-to-end approach where LLMs directly generate masked text with dynamically determined granularity. This resolves the fundamental trade-off in previous methods by allowing the model to decide appropriate masking units (characters, words, or phrases) based on contextual requirements rather than predetermined structural units.

LLM-Assisted Detecting and Redacting Confidential Information for Government Information Disclosure

Masaki Hasegawa

(GENERAL AUDIENCE ABSTRACT)

The rapid growth of generative AI, driven by large language models (LLMs), has led to significant exploration of real-world applications. However, these efforts have largely been led by the private sector, while the adoption of generative AI in public sector organizations, especially in administrative processes, remains limited. This paper explores the "Governmental Information Disclosure Process" as a potential LLM use case for public sector organizations in democratic countries. In many democracies, government agencies are required to disclose information and documents, excluding confidential data, such as personal or sensitive information, upon request from citizens. Typically, administrative bodies must verify and mask confidential content before releasing documents. However, this verification and masking is still done manually, leaving room for efficiency improvements. Moreover, the confidential information to be masked is often defined in natural language, such as legal texts, and requires context interpretation to determine what qualifies as confidential, which is resource-intensive. LLMs, leveraging context and general knowledge, could provide an effective solution. This study evaluates how well LLMs perform in detecting and masking Japanese administrative documents by comparing results from experienced testers, who follow legal guidelines, with those generated by LLMs. This study contributes by proposing end-to-end approach where LLMs directly generate masked text with dynamically determined granularity. This resolves the fundamental trade-off in previous methods by allowing the model to decide appropriate masking units (characters, words, or phrases) based on contextual requirements rather than predetermined structural units.

Acknowledgments

First, I would like to express my deepest gratitude to my advisor, Dr. Shin'ichiro Matsuo, a CCI Research Professor at Virginia Tech's Department of Electrical and Computer Engineering. I first met Dr. Matsuo as an expert in blockchain during my time working for the Japanese government, and he encouraged me to apply to Virginia Tech's Master's program in Computer Science. Dr. Matsuo provided invaluable advice throughout my journey, from course selection to research planning and reporting. His weekly insights on defining problems and conducting quantitative evaluations were a crucial compass for navigating my research.

I also owe sincere thanks to Professor Wenjing Lou, who graciously joined my committee early on and provided invaluable feedback even as my focus shifted from blockchain to AI in administrative efficiency.

I extend my appreciation for Professor Melissa Cameron's thoughtful comments and encouragement on my research progress and schedule. Despite her busy teaching schedule, her support was instrumental, and I am sincerely thankful for her dedication.

I would also like to thank my friends at VT—Vincent, Richard, David, Timo, and Samuel —who made my life colorful and fulfilling. I am equally grateful to the four testers. I also received invaluable advice during my research from Shohei, a researcher at Cyber Smart.

Finally, I express my deepest appreciation to my wife, Reimi, who supported me throughout this journey, and to my parents, who visited twice. Above all, I am forever grateful for my son, Toma, born in the U.S., whose presence brought immense joy and strength. Without their unwavering support, my experience as a graduate student abroad would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The emergence of large language models(LLMs) has triggered significant expansion in generative AI, prompting both academia and industry to invest substantial resources and continue driving its large-scale development. The use cases for generative AI are exploring at an accelerating pace, and recent research has shown that the latest models exhibit high performance even in very advanced reasoning and complex tasks within specific domains.

For example, a report published by McKinsey & Company in 2023 pointed out that generative AI has the potential to significantly impact business functions such as customer operations, marketing and sales, software engineering, and research and development [8]. In line with these forecast and advancement in cutting-edge model development, the use of generative AI and pilot experiments within private companies have been rapidly expanding. As such, the high potential for application in private-sector business domains is becoming increasingly evident, and evaluations of its implementation as well as the accumulation of use cases are progressing at an exceptionally fast pace.

On the other hand, the government is allocating resources to analyze the necessary regulations for AI while prioritizing the assessment of various risks associated with generative AI, such as LLMs. As a result, although there are discussions on expanding use cases that have

proven effective in the private sector, efforts and research analyzing the potential for implementing LLMs in specific public sector tasks or their potential for efficiency improvements are limited [35]. Regarding this situation, the author, who has experience as a civil servant in Japan, speculates that, due to concerns about supply chain risks related to training data and model creation, as well as information leakage, empirical experiments on core tasks have been slow, whether using closed or open-source models of generative AI [46]. On the other hand, there remains considerable inefficiency in current manual workflows in the public sector, and given the significant potential for efficiency gains through the use of generative AI, it is important to discuss the applicability of LLMs to tasks specific to the public sector.

This thesis focuses on the government's response process to information disclosure requests as a specific task in the public sector. A key pillar supporting democracy in many mature democratic countries is the information disclosure system related to government-held information. As detailed in Chapter 3, this process involves a large number of requests, and since verification tasks for confidential information are manually performed, there is a high demand for efficiency improvements. If efficiency can be achieved using LLMs, the potential impact would be significant. Additionally, during the information disclosure process, there is a need to mask confidential information, and since the definition of confidential information is written in natural language legal texts, and the application of these definitions depends on the context of each document, LLMs, which are capable of flexible reasoning based on context, may be well-suited for this task.

We analyze the specific task of government information disclosure process in detail and extract the task of detecting and masking confidential information in documents, which is the process where the potential impact of LLM adoption is most anticipated. One research challenge is the severe lack of studies examining the introduction of LLMs in specific public sector tasks. While there has been research on the use of LLMs for masking PII (Personally

Identifiable Information) in the private sector, as well as studies on detecting personal information contained within LLM prompts, to the best of my knowledge, research applying this to information disclosure in the public sector is limited to studies by Branting et al. and Baron et al., who investigated the effectiveness of LLM-based methods for detecting a limited category of confidential information (deliberative language) in U.S. federal government Freedom of Information Act (FOIA) requests [4, 5]. However, their research only support deliberative language detection at the sentence level in documents and are not well-suited for detecting and masking a wider range of confidential information with finer granularity, such as word-level or character-level, as typically required in actual government information disclosure processes.

1.2 Research Challenges

As mentioned above, detecting and masking confidential information for government information disclosure is a common practice in many democratic nations, with a high volume of disclosure requests. While the definition and categories of confidential information are typically defined by natural language in the law, determining which specific information qualifies as confidential in individual documents requires careful consideration based on the context of the document and other social circumstances. If confidential information is inadvertently disclosed, such as personal information or corporate secrets, the government may face lawsuits or claims for damages. Moreover, if the disclosed information pertains to national security or public safety, it could be exploited, significantly undermining public trust in the government. Therefore, precise and accurate processing is essential.

This research aims to analyze how effectively LLMs can contribute to improving the efficiency and accuracy of the detecting and masking process in government's information disclosure.

However, as stated in Section 1.1, there appears to be very limited prior research exploring the use of LLMs for detecting and masking confidential information in public-sector disclosure processes. In addition, since this study focuses on the information disclosure process in the Japanese government, it was necessary to create a sample dataset similar to Japanese administrative documents from scratch.

As detailed in Chapter 2, existing studies on the use of LLMs for detecting and masking sensitive information in the private sector primarily target PIIs. In contrast, the information disclosure process in the public sector requires targeting a broader range of confidential categories, such as corporate secrets, national security information, and public safety information. To conduct detecting and masking experiments using LLMs, including creating sample documents with such information, establishing evaluation criteria, and executing the experiments, requires both extensive domain knowledge and meticulous effort in constructing sample datasets. These constitute the research challenges addressed in this study.

1.3 Contributions

Zero-shot masking of confidential information using LLMs is a cost-effective approach that does not require additional training, making it promising for improving government information disclosure processes. However, previous approaches face a critical challenge: sentence-level processing fails to provide the fine-grained redactions required in real-world applications, while word-level processing struggles to capture broader contextual dependencies that span multiple sentences. To address this challenge, we propose a novel end-to-end zero-shot approach that fundamentally reframes the task. Rather than using LLMs to make binary redaction decisions, we leverage their generative capabilities to directly produce masked text output. This generative approach enables adaptive masking granularity, capable of masking

anything from entire paragraphs to individual characters as contextually appropriate, while maintaining document-wide contextual awareness. This approach effectively overcomes the rigid granularity constraints inherent in traditional methods.

Our study makes the following contributions:

1. **Adaptive-Granularity End-to-End Masking:**

   We propose an end-to-end zero-shot approach where LLMs directly generate masked text with dynamically determined granularity. This resolves the fundamental trade-off in previous methods by allowing the model to decide appropriate masking units (characters, words, or phrases) based on contextual requirements rather than predetermined structural units.

2. **Category-Aware Framework and Specialized Dataset:**

   We develop an evaluation methodology tailored for generative masking outputs and a data set with confidential information notated by categories at character level. This integrated contribution enables rigorous assessment and counterbalances potential accuracy loss in our end-to-end approach by incorporating confidentiality category awareness into the masking process, enhancing both precision and explainability.

3. **Cross-Cultural and Linguistic Adaptation:**

   We demonstrate our approach's value for Japanese documents, addressing both linguistic challenges (non-segmented text where character-level processing is advantageous) and institutional differences (limited public redaction datasets and distinct disclosure practices). This provides insights into adapting confidentiality masking techniques to non-Western linguistic and regulatory contexts.

1.4 Organization of the Thesis

In Chapter 2, a comprehensive review of the literature is conducted on two main research trends related to the purpose of this research, and the current research findings will be organized. These include works about LLM adoption for the public sector and related works about using LLM for tasks dealing with secret information, such as PII. In addition, in this chapter, we present a limited number of studies on the detection of sensitive information in government information disclosure using LLMs, positioned at the intersection of these two directions, and highlight the differences between these studies and the present research.

Chapter 3 gives an overview of the information disclosure system by administrative agencies, focusing on the case of the major democratic countries. This chapter will also identify areas where modeling the workflow of this study and the introduction of LLM will be considered. Furthermore, the potential needs for introducing LLM in this workflow is discussed.

Chapter 4 provides an overview of the methodology employed in this study. Specifically, it explains the overall process of the research, including dataset creation, system implementation, experimentation, and evaluation of results.

In Chapter 5, we explain in detail the method of creating the custom dataset, including its content. As there were no existing datasets, especially about confidential information detection in Japanese, available for open-source use, this study begins by creating a custom dataset in Japanese.

In Chapter 6, we compare the detecting and masking results of LLM and human testers using the original data set and evaluate how accurately LLM can detect and mask confidential information. This evaluation will consider various variables such as different LLM models, prompt types, and work strategies, and will analyze how to minimize hallucinations and errors during the process of detection and masking. Furthermore, based on the categories

of confidential information, this chapter will assess the current feasibility of using LLM by dividing the evaluation into PII and other categories of confidential information defined by government agencies.

Finally, in Chapter 7, we concludes the thesis by summarizing the experimental results and discussing directions for future research.

Chapter 2

Related Work

In this chapter, we conduct a comprehensive review of two research lines that are closely related to the focus of this thesis: the applicability of LLMs in the public sector and the utilization of LLMs for tasks involving sensitive information, such as PII. First, through a review of existing research, we highlight the limited scope of studies and initiatives regarding the use of LLMs in the public sector, aside from general discussions or cross-sector applications inspired by private sector use cases. Next, by conducting an extensive review of the relationship between LLMs and the handling of sensitive/confidential information, we will clarify the current research landscape on how LLMs are applied to the detecting and masking of sensitive/confidential information. Furthermore, we will organize the distinctions between these studies and the specific focus of this thesis.

2.1 Applicability of LLMs in the Public Sector

When public institutions such as governments engage with AI systems like LLMs, their involvement can be broadly categorized into two scene: the development of regulatory and supervisory frameworks, and the consideration of LLMs for enhancing the efficiency and sophistication of public services and policy-making. Regarding the former, issues such as balancing regulation with innovation and ensuring the effectiveness of global regulations are highly complex and the subject of active debate [11, 13, 28]. However, this research focuses on the latter—examining the use of LLMs in public sector applications.

Although some pilot experiments have been conducted regarding the use of LLMs in the public sector, the current landscape is still dominated by the horizontal application of use cases developed in the private sector. Efforts to explore the application of LLMs to tasks specific to the public sector are still in their early stages.

2.1.1 Research on the use of LLMs to workflows in public sector

Sanjeev et al. provide a comprehensive explanation of the framework for the introduction of LLMs in the public sector, including risk analysis methods, key considerations, architecture, and more [35]. However, while examples are provided for each LLM task, many of these cases are similar to use cases in the private sector, and they do not delve into a specific evaluation of the applicability and potential for use in public sector operations. While the economic and social impacts of LLM systems are receiving significant attention, there is a lack of detailed analyses on the applicability and effects of LLMs in specific public sector operations. Jonathan et al. conducted a survey of 938 public service providers in the UK (including education, healthcare, social welfare, and emergency services) and found that 45% of respondents recognized the use of generative AI in their workplaces, with 22% confirming

the actual use of generative AI systems. Public sector professionals reported positive views regarding current technology use and the potential for future efficiency improvements [6]. Similarly, Fang and Xu evaluated a government-specific LLM-based QA guidance system for Chinese government operations, combining citizen inquiry data and LLM [12]. Their model showed a 4-10% performance improvement over existing models, generating responses that closely resemble natural human language and emphasizing the applicability of LLMs for QA systems in the public sector. Additionally, Vincent et al. analyzed the potential for generative AI to reduce costs and bureaucratic overhead in government services [38]. They assessed the scale of bureaucratic decision-making procedures for citizen services in the UK central government and measured the potential for automation via generative AI. Their findings indicated high potential for labor savings and pointed out that automation efforts should focus on general procedures, rather than the services themselves. Viechnicki and Eggers' research revealed that the average public servant spends up to 30% of their time on information recording and other basic administrative tasks, further supporting the effectiveness of generative AI [45]. Looking at the parliament, Lucke and Frank investigated the potential use of generative AI in parliamentary routines [23]. They concluded that, for multifaceted questions, generative AI could provide content for decisions, documentation, and meeting preparations, but emphasized that the generated content should be used as drafts for further review by humans.

2.1.2 Pilot Experiments Utilizing LLMs in Government

As mentioned in Subsection 2.1.1, while the accumulation of research cases on the utilization of LLMs in the public and government sectors is still in its early stages, several countries have started pilot projects to explore the use of LLMs in government operations. For instance, the UK government began an experiment with a generative AI chatbot, "GOV.YK Chat,"

in 2023 [41]. This tool allows small business owners to ask questions about government information and support programs and was developed using OpenAI's GPT-4. The chatbot aims to quickly navigate users through complex and dispersed government information and provide individualized, integrated responses. While the experiment is still in its early stages, with a limited number of participants, nearly 70% of the users reported that the tool was helpful, suggesting potential for future applications. However, issues such as the accuracy and correctness of responses, as well as handling questions it should not answer, have also been identified. Additionally, Japan's Digital Agency conducted technical verification for the appropriate use of generative AI in administration from 2023 to 2024 [10]. Over 10 government agencies and more than 20 local governments participated in this experiment, which involved identifying services in administrative tasks that could be improved by generative AI and estimating the impact on operational improvements. The applicability of generative AI to tasks such as automatic document creation, response creation for inquiries, document proofreading, drafting responses, creating technical documents like programming code, and determining the confidentiality level of documents, which partially align with the scope of this study, was tested, showing certain effectiveness. Around 90% of participants reported improvements in work efficiency and the quality of deliverables with the use of generative AI. Furthermore, the U.S. Department of Homeland Security (DHS) has released an "AI Use Case Inventory" summarizing the current usage of AI [43]. Among the use cases related to generative AI or LLMs, Homeland Security Investigations (HSI) is conducting tests with an LLM-based system that leverages open-source technology to help investigators more quickly summarize and search for contextually relevant information within investigation reports [9]. This system is expected to improve detection of drug-related networks and the accuracy of identifying crimes such as auto-extraction fraud. Additionally, the United States Citizenship and Immigration Services (USCIS) has launched a pilot project to improve the training of immigration officers using an interactive application that leverages GenAI [9].

2.2 Utilization of LLMs for Tasks Involving Sensitive Information

The other research area closely related to this study is the work of handling sensitive information, such as personal information identifier(PII), in processes using LLMs. This research vector can be broadly divided into two directions. The first focuses on detecting and masking PII from documents in the private sector. This line of research shares many similarities with the present study, as both use LLMs to detect confidential information within documents. However, a key distinction is that prior studies have focused mainly on personal information, whereas this study addresses a broader range of confidential categories in governments. The second research vector seeks to detect and replace inappropriate or privacy-sensitive information in the input and output of LLMs. This research aims to implement safeguards that ensure that LLM inputs and outputs do not contain problematic content. Although the actual approaches of both research vectors are very similar, they differ in their intended tasks. Therefore, this section reviews them separately.

2.2.1 Detection of PII in Documents Using LLMs

There are several studies that aim to utilize LLM-based approaches to detect and sanitize confidential information in documents within the private sector [2, 14, 22, 27, 36, 37, 48].

Yang et al. evaluated the performance of LLMs in the task of detecting PII within archival data in Chinese organizations [48]. Their results show that even when only prompt-based methods were used, LLMs exhibited practical performance in detecting PII within large amounts of archival data, suggesting the potential of using LLMs to identify data that contain personal information from large data sets. Federico et al. proposed a text sanitiza-

tion method using pre-trained LLMs without any additional fine-tuning [2]. Their method detects and replaces potentially sensitive words, specifically focusing on PII, within given unstructured text data. The advantages of their LLM-based method include its ability to handle unstructured data such as dialogues, its domain-independent nature without requiring additional training, and its capacity to maintain data utility by replacing sensitive terms with contextually consistent alternatives generated by LLMs. Staufer et al. pointed out the necessity of sanitizing not only PII but also information that could indirectly lead to individual identification in whistleblowing scenarios [37]. They proposed a method to reduce identification risks even further by using LLMs to process text while maintaining the context of the original whistleblowing text. Their study demonstrated that combining existing anonymization methods with LLMs significantly reduces authorship identification accuracy while preserving the context of the original text. Similarly, Robin et al. highlighted the need for anonymization against adversarial inferences of personal data from online text using LLMs and developed an LLM-based adversarial anonymization framework to counter such attacks. According to their experiments, LLM-based anonymization outperforms existing anonymization tools in terms of balancing privacy protection and utility [36].

In the medical domain, where it is crucial to sanitize sensitive information such as patient data, there has been extensive research on text sanitization. García et al. conducted experiments on anonymizing personal information in large-scale Spanish medical data using BERT-based models [14]. Their performance evaluation revealed that these models achieved high performance without any domain-specific training, outperforming traditional anonymization methods. Li et al. further investigated the utility of larger models and confirmed their effectiveness in anonymizing medical data sets [22]. They developed a framework for anonymizing medical data by incorporating ChatGPT and GPT-4, demonstrating their superior capabilities in masking sensitive information. However, they also pointed out challenges such as

the lack of transparency in closed LLM implementations and the risks of handling sensitive data through online APIs. These challenges highlight the importance of developing high-performance models that can be deployed in local environments in the medical domain.

Mori et al. proposed a method using GPT-4 to identify sentences containing confidential information in corporate documents [27]. Although their experiments were limited to a single document type, in this case academic paper in English, their study demonstrates the potential of using LLMs to improve the efficiency of detecting and managing confidential information in corporate documents at the sentence level.

2.2.2 Detecting Problematic contents in LLM's Input and Output

Research on guardrails to detect privacy-related or confidential information in the input of LLMs, prevent adversarial inputs, and prevent the generation of high-risk or policy-violating content in the output has been actively conducted, not limited to sensitive information. Based on these studies, systems for inspecting the output content and detecting and removing harmful content or policy violations have been provided [21, 24, 32]. In addition, Inan et al. at Meta developed Llama Guard, an LLM-based input-output safeguard model applicable to human-AI conversations [16]. Their model, which can be used as a separate LLM to perform classification tasks based on commonly considered risk categories related to AI inputs and outputs, demonstrated performance surpassing that of other existing content inspection tools.

Additionally, focusing on PII that may be included in prompts, Sun et al. proposed De-Prompt, a desensitization protection and effectiveness evaluation framework for prompts [40]. Their approach recognizes the differential impact of various entities in prompts on both semantics and privacy, leveraging the robust text comprehension and generation capabilities

of LLMs to achieve prompt sanitization that is resilient to adversarial generation. Similarly, Chong et al. proposed Casper, a technique for sanitizing user privacy information contained in prompts in the context of using web-based LLM services [7]. Casper functions locally as a browser extension, detecting and removing sensitive information such as PII from user inputs before sending them to LLM services, achieving high accuracy in tests.

2.3 LLM-Assisted Government Information Disclosure

Research on utilizing LLMs for detecting confidential information in the information disclosure process within the public sector remains limited. Branting et al. investigated the performance of machine learning techniques in detecting deliberative language, a category of sensitive information in the U.S. federal government's information disclosure process [5]. They examined the detection accuracy of deliberative language using both traditional machine learning classification models and a BERT model-trained on a corpus annotated at the paragraph level—derived from the Clinton White House Collection curated by Barton et al. [3]. Their findings indicate that the BERT model, an LLM trained on their custom corpus, generally outperformed traditional machine learning classification models in this task. Based on these results, they implemented the FOIA Assistant, which supports the detection of confidential information in the government administrative document disclosure process by employing the BERT model for deliberative language and an existing classifier for PII. This system has undergone operational testing across multiple federal agencies.

Furthermore, Baron et al. conducted a preliminary experiment to assess the feasibility of using ChatGPT for the task of detecting deliberative language with a same dataset [4]. Their study is considered the first to apply a chat-based LLM to confidential information detection in government information disclosure. However, their research was limited to only a subset

of confidential information categories, and detection was performed at the sentence level. Their experimental results showed no significant difference in simple detection accuracy between ChatGPT-3.5 and existing classifiers. However, they also highlighted the potential of ChatGPT-3.5 to assist humans by providing justifications for non-disclosure decisions.

These two studies represent early attempts at improving the efficiency of confidential information detection and masking in government information disclosure using LLMs. However, several challenges remain to be addressed. For instance, in government information disclosure, the granularity of masking is typically required to be as fine as possible—often at the character or word level—whereas existing studies have only performed sentence-level detection or masking. Additionally, in both studies, BERT and ChatGPT were fed sentence-level inputs, which prevented the models from incorporating contextual dependencies across multiple paragraphs or entire documents when determining confidentiality.

To address these challenges, this study adopts an end-to-end approach in which larger textual units, such as entire paragraphs or document files, are inputted to LLMs, and the output consists of text where confidential information is replaced with masking symbols. This approach enables an evaluation of LLM effectiveness in a setting more aligned with real-world applications. Moreover, while existing studies have focused on dealing with only a single category of confidential information, this study leverages the ability of chat-based LLMs to perform various tasks in a zero-shot setting without requiring training. Specifically, it conducts detection and masking task covering multiple legally defined confidential information categories. Thus, while building upon prior research in this domain, this study introduces novelty by empirically examining the effectiveness of LLMs in a more practical masking approach.

Additionally, it focuses on the government document disclosure process in Japan, investigating the potential for workflow improvement through LLM applications. To achieve

this, we construct a Japanese-language dataset for empirical evaluation. To the best of our knowledge, no prior studies have attempted specific masking of sensitive information using Japanese text, further underscoring the novelty of this research.

Chapter 3

Model of Information Disclosure System in the Government

In this chapter, we summarize the process of information disclosure requests within government, which is the focus of efficiency improvement using LLMs in this study, along with an overview of information disclosure systems in major democratic nations. After highlighting the necessity of efficiency improvements of the process, we focus on the case of the Japanese government and conduct modeling for this research.

3.1 Overview of Information Disclosure System of Governments in the World

Information disclosure system by government serves as a fundamental institution of democratic country, and similar information disclosure systems have been enacted as legal frameworks in many democracies. As shown in Table 3.1, these systems generally grant citizens the right to request the disclosure of any information from the government. With a few exceptions, governments are required to disclose requested information within a specified period in principle. Additionally, all countries define specific categories of nondisclosable information, such as personal privacy data and information related to national security, through legal provisions. These conditions are described in natural language within the respective legislation. Furthermore, the laws underpinning these systems in each country typically mandate that decisions regarding the disclosure or non-disclosure of information must be processed within approximately one month. Enhancing the efficiency of these processes within governments can have a significant impact. Notably, even in cases involving requests for large volumes of information, the standard processing period remains unchanged in principle.

Although there are slight differences in the systems of each country, the general information disclosure process is as shown in Figure 3.1.

- **Step 1**: A request for information disclosure from citizens is received by the government agency.
- **Step 2**: The government investigates whether the requested information is held.
- **Step 3**: If the information is held, it is examined to determine whether its disclosure would involve disclosing confidential information, and any confidential information is removed.

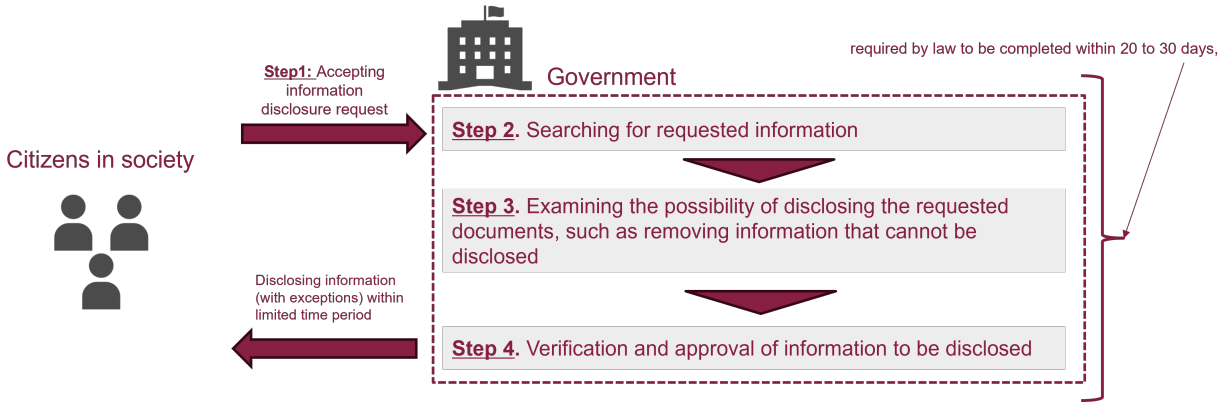

Figure 3.1: General Procedure of Information Disclosure by Government

- **Step 4**: The government agency makes a decision within the organization and takes action, such as disclosing or withholding the document, or not responding about the existence of the information, and communicates the result to the citizen.

Based on these steps, Section 3.4 will model the specific process in Japan, using this as a reference for this study.

An overview of the systems in the United States, the United Kingdom, and Japan will be provided in the following subsections.

Table 3.1: Comparison of information disclosure systems in major countries.

Country	**Legal basis**	**Description**	**Exception**	**Deadline**
U.S.A.	Freedom of Information Act (FOIA)	Citizens have the right to request the disclosure of information from central government agencies and public institutions. The government adheres to the principle of information disclosure.	Information related to national security or personal privacy, etc.	20 BD
U.K.	Freedom of Information Act 2000	Citizens have the right to request the disclosure of information from central government agencies and public institutions. The government follows the principle of information disclosure.	Information related to national secrets or personal privacy, etc.	20 BD
Japan	Act on Access to Information Held by Administrative Organs	Citizens are granted the right to request the disclosure of information held by central and local governments.	National secrets and personal information, etc.	30 days
Canada	Access to Information Act	It mandates the disclosure of federal government information to citizens.	Information related to national security or personal privacy, etc.	30 days
Australia	Freedom of Information Act 1982	Citizens have access to federal government information.	Information related to national security or personal privacy, etc.	30 days
Germany	Informations freiheitsgesetz (IFG)	Citizens have the right to request the disclosure of information from federal government agencies.	Disclosure may be restricted for reasons such as privacy or national security.	1 month

3.1.1 the United States: Freedom of Information Act (FOIA)

The Freedom of Information Act (FOIA), enacted in 1966, is a federal law that requires the disclosure of all or part of the information managed by the U.S. government [42]. This law is described on the U.S. government's webpage as a law that enables citizens to learn about their government and is considered an essential part of democracy. Under this law, federal agencies are mandated to disclose information requested under FOIA unless it falls under one of the nine exemptions listed in Table 3.2, which protect interests such as personal privacy, national security, and law enforcement.

FOIA also specifies that information should be withheld only if its disclosure would reasonably be expected to harm the interests protected by the exemptions, or if disclosure is prohibited by law. This means that federal agencies are generally required to disclose information unless it pertains to classified material. Additionally, FOIA requires that if full disclosure is deemed impossible, agencies must consider partial disclosure by segregating non-exempt information and taking reasonable steps to release it. This approach is similar to typical government information disclosure systems in other democratic nations.

Citizens must submit a FOIA request to the FOIA office of the relevant government agency. Each agency is required to comply with FOIA and adhere to government-wide guidance created by the Department of Justice's Office of Information Policy. While FOIA mandates a response to requests within 20 business days, an investigation of the 2023 FOIA Annual Reports revealed that, among the 122 agencies with organized data, over 50 agencies exceeded the average processing time of 20 days. In some cases, the average processing time exceeded 200 days. Even when considering median response times, more than 20 agencies failed to meet the statutory response deadline. This suggests that screening requests for classified or sensitive information is challenging and time-consuming for many government agencies [44].

Table 3.2: Exemption of FOIA [42]

Exemption	Description
Exemption 1	Protects information that is properly classified under criteria established by an Executive Order to be kept secret in the interest of national defense or foreign policy.
Exemption 2	Protects information related solely to the internal personnel rules and practices of an agency.
Exemption 3	Protects information specifically exempted from disclosure by another statute, if that statute either: (1) requires that the matters be withheld from the public in such a manner as to leave no discretion on the issue; or (2) establishes particular criteria for withholding or refers to particular types of matters to be withheld. An Exemption 3 statute must also cite specifically to subsection (b)(3) of the FOIA if enacted after October 28, 2009.
Exemption 4	Protects trade secrets and commercial or financial information that is obtained from outside the government and that is privileged or confidential.
Exemption 5	Protects certain records exchanged within or between agencies that are normally privileged in the civil discovery context.
Exemption 6	Protects information about individuals in personnel and medical files and similar files when the disclosure of that information would constitute a clearly unwarranted invasion of personal privacy.
Exemption 7	Protects records or information compiled for law enforcement purposes, but only in cases where the disclosure of such law enforcement records or information would cause specific adverse effects.
Exemption 8	Protects information contained in or related to examination, operating, or condition reports prepared by, on behalf of, or for the use of, an agency responsible for the regulation or supervision of financial institutions.
Exemption 9	Protects geological and geophysical information and data, including maps, concerning wells.

3.1.2 the United Kingdom: Freedom of Information Act 2000

The Freedom of Information Act 2000 is a law enacted by the UK Parliament that establishes public access to information held by public authorities in the UK [30, 31]. Under this law, requests for information can be made to government departments, delegated administrative bodies, local councils, schools, NHS-related entities, and public authorities such as police and fire departments. Citizens can submit Freedom of Information requests in writing, including online submissions, to the relevant public authority. Additionally, each public authority

publishes disclosure logs online, allowing access to previously disclosed responses to past FOI requests. Legally, public organizations that receive requests are required to disclose the information within 20 working days. Furthermore, Sections 21-44 of the Act outline exemptions from disclosure, which are summarized in Table 3.3. Further detailed explanations of each exemption can be found on the UK government's Information Commissioner's Office website [17]. The types of these exceptions have different legal structures, but fundamentally, they align with the approach to exceptions in the U.S. and Japan. Furthermore, by the Act, the public organization receiving a request is required to disclose the information within 20 business days. According to the Information Commissioner's Office website, in certain cases, the public organization can extend the deadline in accordance with the law. However, as a best practice, Section 45 of the Act stipulates that a 20-day extension is considered the best practice. For instance, if the requested information is thought to fall under an eligible publication exemption, an extension may be granted to conduct a statutory public interest test [18].

Table 3.3: Exemptions under the Freedom of Information Act (FOIA) - UK

Exemption	**Description**
Section 21	Information already reasonably accessible
Section 22	Information intended for future publication
Section 22A	Research information
Sections 23 and 24	Security bodies and national security
Sections 26 to 29	(Various exemptions related to law enforcement and security)
Sections 30 and 31	Investigations and prejudice to law enforcement
Section 32	Court records
Section 33	Prejudice to audit functions
Section 34	Parliamentary privilege
Sections 35 and 36	Government policy and prejudice to the effective conduct of public affairs
Section 37	Communications with the royal family and the granting of honours
Section 38	Endangering health and safety
Section 39	Environmental information
Section 40(1)	Personal information of the requester
Section 40(2)	Personal information
Section 41	Confidentiality
Section 42	Legal professional privilege
Section 43	Trade secrets and prejudice to commercial interests
Section 44	Prohibitions on disclosure

3.1.3 Japan: Act on Access to Information Held by Administrative Organs (AAIHAO)

The Act on Access to Information Held by Administrative Organs(AAIHAO) is a Japanese law enacted in 1999 that governs the procedures for making requests for the disclosure of information held by administrative organs of the Japanese government [19]. Based on the principle of popular sovereignty, the law aims to promote a fair and democratic administration under the scrutiny and criticism of the public by regulating the right to request the disclosure of administrative documents held by national administrative organs. The law obligates national administrative organs to disclose information, but does not apply to local governments, courts, or the National Diet. The law defines the scope of information that can be disclosed as "documents, photographs, or electronic records created or obtained by the staff of administrative organs in the course of performing their duties, and held by the administrative organs for the organizational use of their staff." With the exception of non-disclosure information listed in Article 5 (see Table 3.4), the government is required to disclose the information it holds. Furthermore, when a request for disclosure is made for an administrative document that contains non-disclosure information, the remaining portion must generally be disclosed by removing or masking the non-disclosure information. Administrative organs are legally required to notify whether or not to disclose the requested document within 30 days from the day the request is made. However, if there is an unavoidable reason, such as a large volume of requested documents, the deadline may be extended under certain conditions.

Table 3.4: Non-Disclosure Information under AAIHA

Exemption	Description
Information concerning an individual	• Information that identifies an individual through name, date of birth, etc. • Information that cannot identify an individual, but disclosure could harm the individual's rights and interests.
Information concerning a juridical person or other organizations	• Information likely to harm the rights, competitive position, or legitimate interests of a juridical person, etc. • Information voluntarily provided under non-disclosure conditions by administrative organs.
National security-related information	• Information that may harm national security, damage international relations, or create disadvantages in negotiations.
Crime prevention or public safety-related information	• Information that could interfere with crime prevention, investigations, prosecutions, or the maintenance of public safety.
Government deliberations and consultations	• Information that may harm the neutrality of decision-making or cause unfair benefits or disadvantages in consultations.
Government affairs or business	• Information that may hinder accurate fact-finding related to audits, inspections, supervision, examinations, or tax matters. • Information that may facilitate wrongful acts or make it difficult to discover such acts. • Information that may unjustly harm the financial interests or position of the government, public entities, or local authorities. • Information that may unjustly hinder fair and efficient execution of research and study. • Information that may hinder the fair and smooth management of personnel matters. • Information related to the business of state-run enterprises or local public organizations that could harm their legitimate interests.

3.2 The Need for Process Efficiency: A Focus on Japan

As summarized in Table 3.1 and Section 3.1, the legislative frameworks concerning government information disclosure obligations in major democratic nations are similar. Citizens are granted the right to access and request disclosure of government-held information, and governments are obligated to disclose such information unless it falls under limited exceptions defined by law. This study focuses on Japan as a case study among major democratic countries. Given the similarities across advanced nations, modeling and experimentation based on Japan does not compromise generalizability.

The statistics of Information Disclosure requests in Japan is summarized in Table 3.5. Furthermore, the number and the proportion of cases with extension deadline for processing is illustrated in Figure 3.2 [25]. The data reveals that approximately 200,000 requests are filed annually across all government agencies in Japan, equivalent to about 2% of the population submitting at least one request per year. Considering that each request may target multiple documents containing specific information, the total number of documents requested could be several times higher than the number of applications.

The processing of these requests is carried out by individual government agencies. As discussed in Section 3.1, the workload of investigating document existence, verifying confidentiality, and redacting sensitive information places a significant operational burden on agencies. As shown in Figure 3.2, approximately 10% of annual information disclosure requests exceed the statutory processing deadlines due to heavy workloads and require deadline extensions. With over 10,000 cases failing to meet statutory deadlines annually, the necessity for efficiency improvements in government operations is evident.

Additionally, this challenge is not unique to Japan. As highlighted in the latter part of Subsection 3.1.1, similar difficulties are observed in the United States, where many requests fail

to be processed within the statutory 20-business-day deadline. This suggests that managing the burden of information disclosure requests is a common global issue.

Table 3.5: Information Request Processing Data in Japan [25]

Number	**FY2019**	**FY2020**	**FY2021**	**FY2022**	**FY2023**
Annual requests	160,546	164,950	178,386	185,673	192,576
On time (30BD)	146,538	147,094	160,763	169,431	175,461
The cases with deadline extensions	14,008	17,856	17,623	16,242	17,115
Extension ratio(%)	8.73	10.83	9.88	8.75	8.89

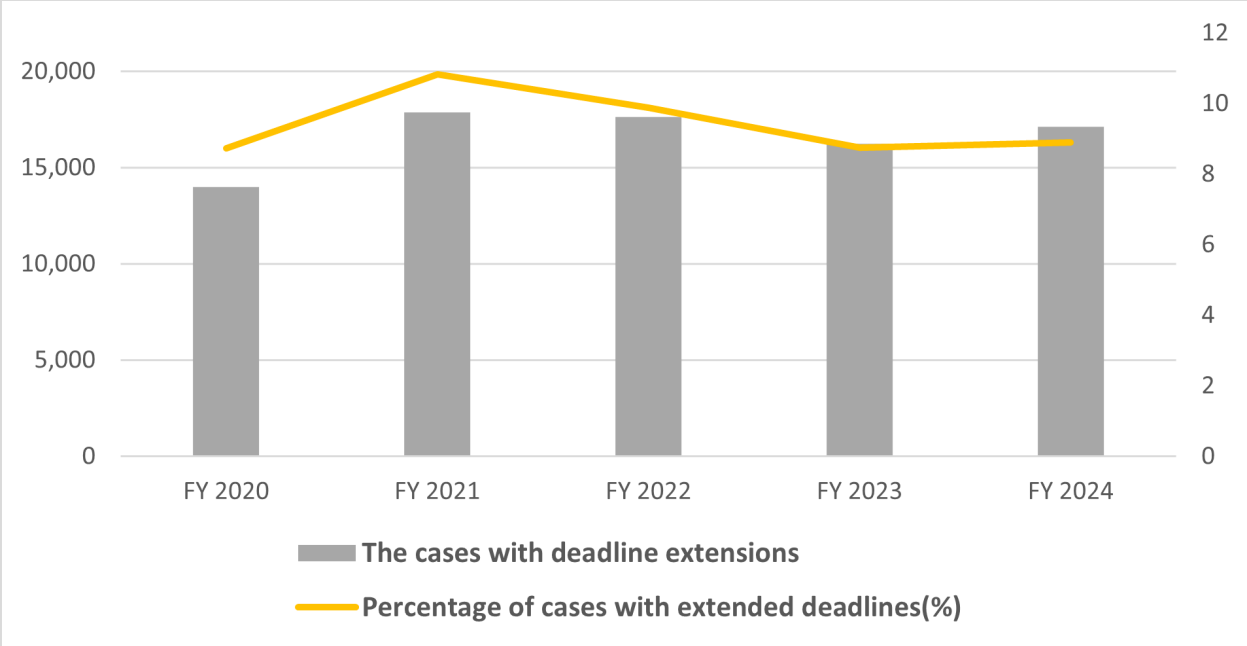

Figure 3.2: The number of deadline extensions and their proportion of the total requests (FY2019-23)

3.3 Large Language Models and Its Applicability

While there is no globally agreed-upon strict definition of LLMs, they are often described as neural network-based natural language models capable of performing various tasks through

pre-training on massive datasets and an enormous number of parameters that estimate the probabilistic distribution of text [29, 47]. As noted in the research by Yang et al., LLMs acquire highly generalized capabilities in context comprehension through extensive training on large datasets, enabling them to make context-aware judgments in identifying confidential information such as PII [48]. For instance, they may be effective in detecting information that should be classified as confidential due to its association with other data—an area where traditional rule-based detection methods have faced difficulties. In this study, we explore the potential of leveraging LLMs' advanced contextual understanding abilities for detecting categories of confidential information beyond PII, aiming to enhance the efficiency of governmental information disclosure processes.

3.4 Modeled Flow of Information disclosure Process

In this section, we model the workflow for processing information disclosure requests within the Japanese government and explain where in the workflow the detection and masking process for confidential information, a key focus of this study using LLMs, is positioned.

Figure 3.1 provides an overview of the general processing steps for information disclosure requests globally. However, the workflow model specific to Japan includes additional processes, as shown in Figure 3.3. This model serves as the basis for the tasks studied in this research.

Below is a summary of each process, with the study focusing on the efficiency improvements in Process 3 using LLMs:

- **Process 1: Verify whether disclosing the existence or nonexistence of documents related to the request could lead to the exposure of confidential**

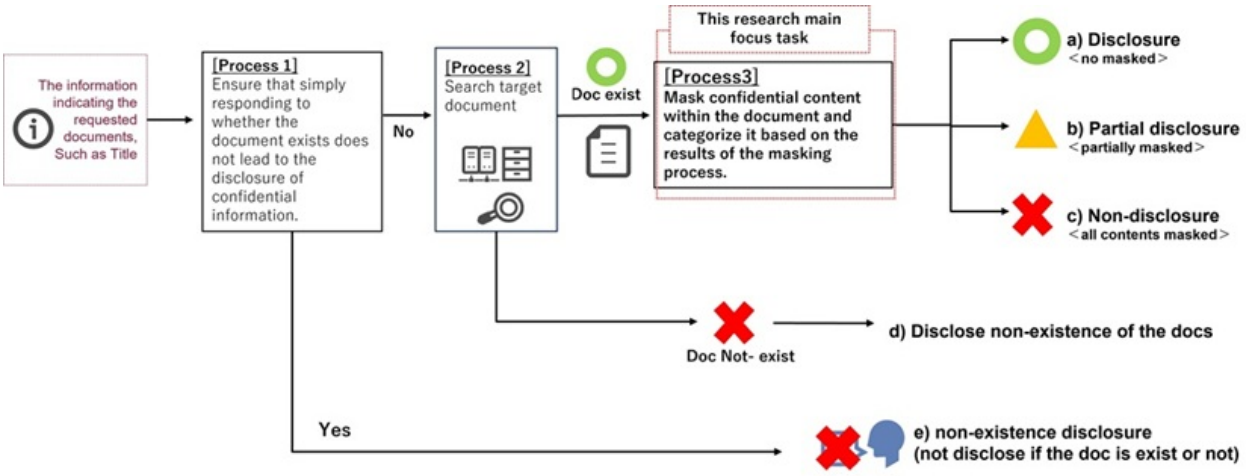

Figure 3.3: Modeling the processing flow of information disclosure requests in Japan

information

Under Article 8 of the AAIHA, it is stipulated that: "When responding to a disclosure request, if merely answering whether or not the administrative documents related to the request exist would result in the disclosure of non-disclosable information, the head of the administrative agency may refuse the disclosure request without revealing whether such documents exist." The government may therefore reject an information disclosure request from a citizen without confirming whether the documents are held, when acknowledging their existence would effectively disclose non-disclosable information. This corresponds to pattern (e) shown in Figure 3.3.

For instance, if information related to meetings between a specific individual and a government agency are requested, answering whether such documents exist would effectively reveal whether the individual had meetings with the agency. This would be considered equivalent to disclosing non-disclosable information about that individual. In this way, even the act of revealing the existence or non-existence of information within a government agency may be deemed equivalent to the disclosure of certain non-disclosable information. Such requests are therefore carefully reviewed before the

documents are searched, and responses are provided in the form of "neither confirm nor deny of existence" replies.

- **Process 2:Search for documents or data related to the requested information.**

  In the verification of Process 1, requests deemed not to fall under "neither confirm nor deny of existence" response, the existence of administrative documents or data related to the requested information within the government agency that received the request will be investigated.

  In the experiment of this study, for simplification purposes, the focus is limited to text-based data. However, in reality, all types of data, including audio, images, and videos held by the government, are subject to information disclosure requests. As such, the government agency that receives a request must investigate whether any administrative documents or data related to the requested information exist in its management systems, such as staff PCs, servers, and physical storage spaces. This investigation also imposes a burden on the administrative agency, particularly when determining whether various types of administrative documents correspond to the requested information. In such cases, a detailed review of the document content is required, making the task very time-consuming. While there is a significant demand for improving the efficiency of this process, it is not the focus of this study. The potential for automating this process using technologies such as LLMs will be a challenge for future research.

  As a result of this process, if the requested information is found, the process will move to Process 3. If the existence of the requested information cannot be confirmed, the response will follow Pattern (d) in Figure 3.3, where the requester is informed that the government does not hold the requested information.

- **Process 3:Examine the discovered documents for confidential information and apply masking to any confidential content**

  Finally, in the investigation of Process 2, if an document containing the requested information is found, the government agency will verify whether the document contains any non-disclosure information as exempted in Article 5 of the AAIHA. If non-disclosure information is found, and it is part of the document (typically at the character level), that part or character will be masked. The result of this process will be classified into patterns (a) to (c) in Figure 3.3, depending on the amount of non-disclosure information contained in the document. This task is currently almost entirely manual, with staff members from the department holding the document and the management department collaborating to carry it out. However, it can become a highly burdensome task, especially when the number of pages of administrative documents that need masking is large. As a result, it is inferred that the processing of information disclosure requests by government agencies often exceeds the specified time frame, leading to numerous extension requests.

  This study will assess the potential use of LLMs for the detection and masking of non-disclosure information in this process.

Chapter 4

Methodology of the Research

4.1 Research Objective

In this chapter, we outline the methodology of this study to evaluate the applicability of LLMs to the task of detecting and masking confidential information in Japanese government-related documents. The primary objective of this research is to examine the technical feasibility and accuracy of detecting and masking confidential information using LLMs, with Japanese documents as a case study. Through experiments, we identify challenges related to hallucinations and detecting and masking accuracy.

4.2 Overview of Methodology

The analysis process in this study consists of dataset creation, system implementation, experimentation, and result evaluation. In particular, the primary components include dataset creation, the implementation of a program for identifying and masking confidential information using LLMs, and the evaluation of results, as illustrated in Figure 4.1.

In summary, this study constructs an original dataset to assess the accuracy of LLMs in detecting and masking confidential information in Japanese text compared to human testers with the professional experiences the in Japanese Government. The dataset comprises orig-

inal texts that may contain confidential information and corresponding ground-truth tags, which indicate the locations and category of such information as identified by human testers. The detection or masking results generated by the LLM are then compared against this ground-truth data to evaluate performance.

The details of each component are provided as follows: dataset generation is discussed in Chapter 5, program implementation using LLMs is explained in Section 6.1, and the evaluation of results is presented in Sections 6.2, 6.3, 6.4, and 6.5.

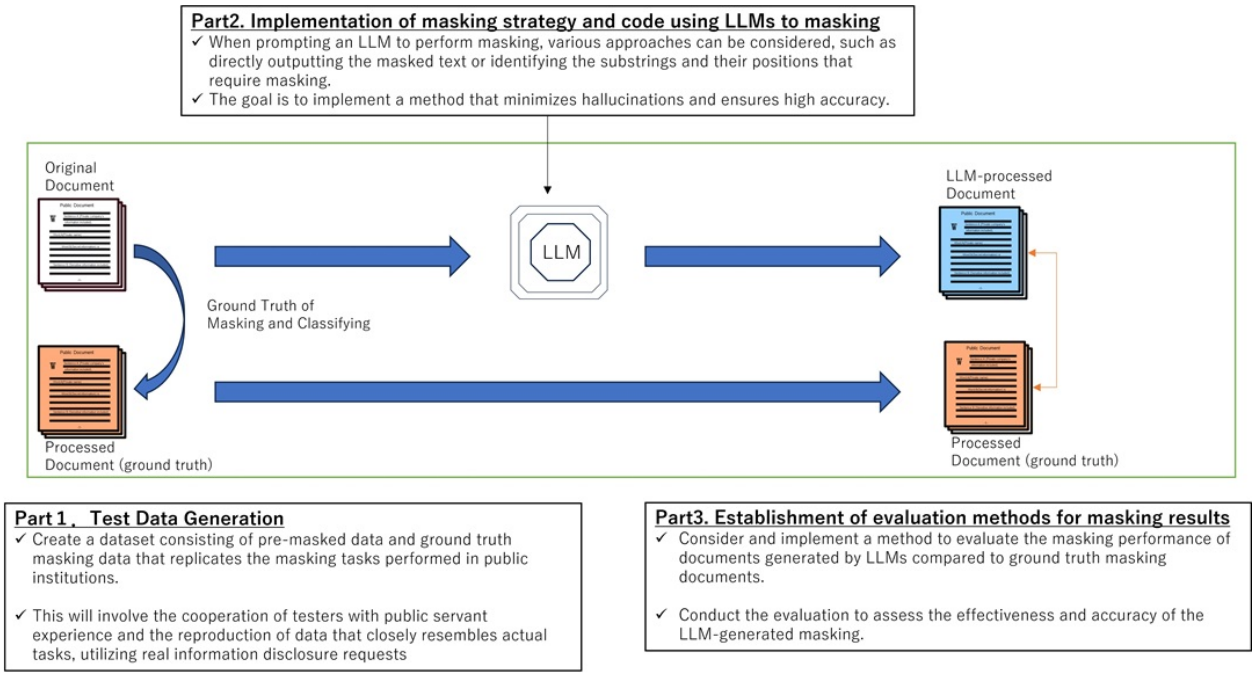

Figure 4.1: Major challenges of the research

Chapter 5

Dataset Generation

In this study, we created a unique test dataset consisting of sample texts containing confidential information in Japanese and their corresponding ground-truth datasets. This chapter first explains the limitations of existing data for such texts and the necessity of creating proprietary data, followed by a detailed explanation of the methodology used to construct the original data set.

5.1 The Need for Developing Original Datasets

In this section, we outline the limitations of existing datasets for detecting confidential information in documents and the challenges associated with accessing government documents in particular confidential. Then we emphasize the necessity of creating an original dataset for this study.

This research evaluates the LLMs' availability for detecting and masking confidential information in documents held by government agencies. As summarized in Chapter 2, prior studies have primarily focused on detecting PII in the private sector. However, there have been limited experimental studies addressing the detection and masking of a broader range of confidential information—such as national security data or corporate secrets—in the context of government information disclosure requests. Additionally, many open-source test datasets for confidential information detection tasks are in English or are limited to PII, making them unsuitable as sample Japanese documents containing the types of confidential information targeted in this study. Furthermore, even when using information disclosure requests, portions of government documents classified as confidential remain inaccessible.

5.1.1 Limitation of Existing Open source Datasets

As summarized in Chapter 2, in the majority of existing studies, those that have experimentally used LLMs for detecting or masking confidential information typically employ test datasets specifically designed for PII (Personally Identifiable Information) detection. [14, 22, 37, 48] Additionally, many other open source datasets have been created for confidential information detection tasks to evaluate PII detection tasks, and both the original and ground-truth text data are limited to PII detection [1, 20, 34]. Since these datasets are in English and do not include any confidential information other than PII, they cannot be

directly used in this study.

5.1.2 Challenges in Accessing Government Documents Containing Confidential Information

Next, since this study uses Japan's government information disclosure requests as a case study, we considered whether it would be possible to use government-held administrative documents directly as experimental datasets. As summarized in Chapter 3, while citizens can request disclosure of information held by the government, any confidential information contained in the disclosed administrative documents will be either masked or withheld. Therefore, as shown in Figure 5.1, document samples that do not contain confidential information can be created based on administrative documents obtained through information disclosure requests from the Japanese government. However, since documents containing confidential information are not disclosed, it was concluded that these document samples must be independently created for the experiment.

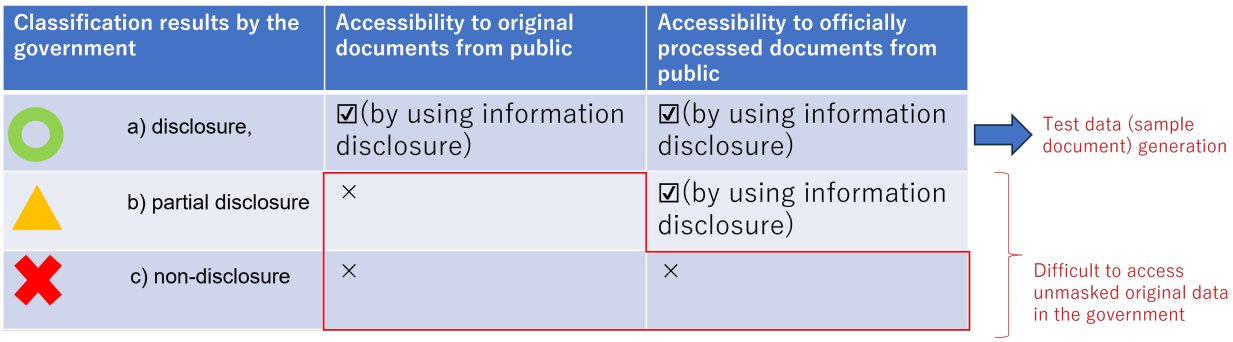

Figure 5.1: Accessibility to Government Documents

5.2 Generation of Pre-Masking Text Data

As stated at the beginning of this chapter, it is necessary to create a demo test dataset consisting of Japanese sample texts containing confidential information and their corresponding ground-truth dataset. The sample texts for detecting or masking confidential information should ideally consist of both positive samples, which contain confidential information, and negative samples, which do not. Ultimately, whether each sample text contains confidential information will be determined based on the results of the masking process conducted by human testers.

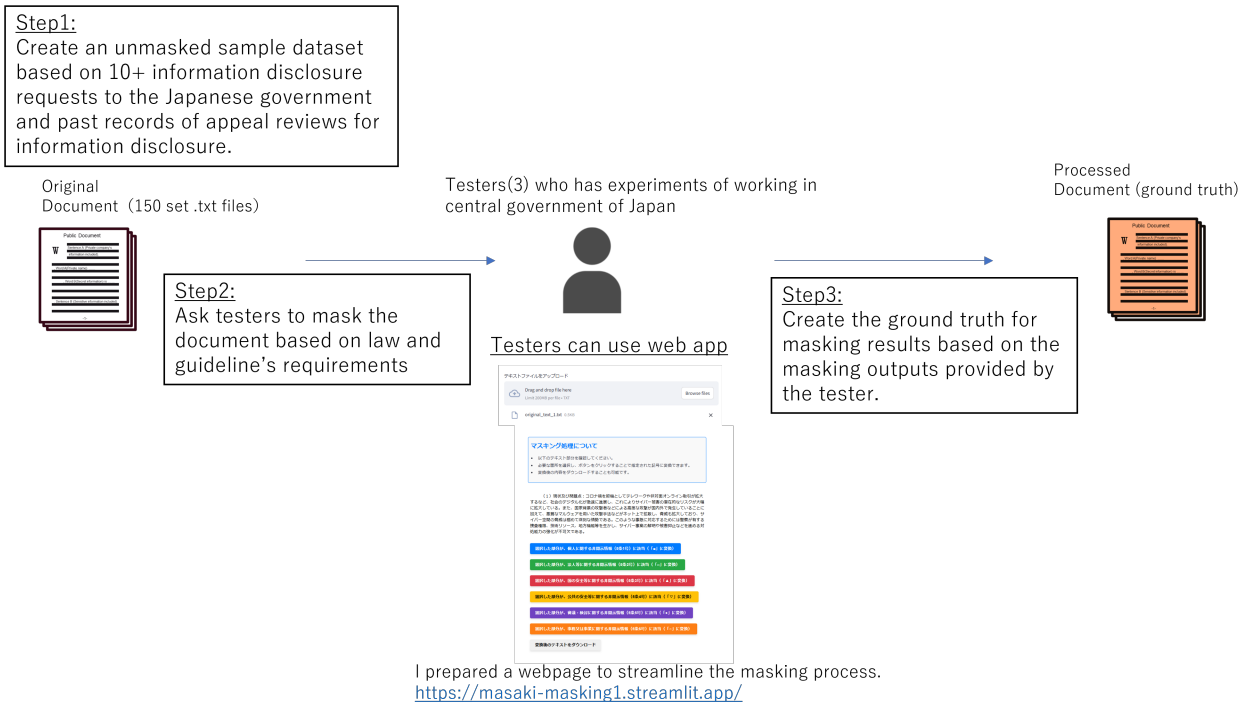

Figure 5.2: Demo dataset generation method

As discussed in Subsection 5.1, the candidate texts for negative samples were created based on documents disclosed without masking through more than ten information disclosure requests submitted by the authors to the Government of Japan. Meanwhile, the candidate texts for positive samples were created by combining the results of more than ten infor-

mation disclosure requests, in which confidential information was partially redacted, with information obtained from the Government of Japan's *Information Disclosure and Personal Information Protection Advisory and Judicial Precedent Database* [26], which provides insights into the handling of appeals regarding disclosure requests. In Step 1, a total of 150 textual files were created.

Furthermore, the types and proportions of confidential information included in the sample texts were referenced from the actual distribution of confidentiality categories in the results of information disclosure requests processed by the Japanese government (see Table 5.1).

Table 5.1: Confidential Categories and Number of Cases in 2023 [25]

Confidential Category	**Number of Cases (Percentage)**
Personal Information	140,783 (86.8%)
Corporate or Organizational Information	129,663 (79.9%)
National Security Information	1,671 (1.0%)
Public Safety Information	5,009 (3.1%)
Deliberation and Review Information	1,678 (1.0%)
Administrative or Operational Information	10,381 (6.4%)

*A single case or document may fall under multiple categories; thus, the total does not sum to 100%.

5.3 Generation of Masked Ground Truth Data

5.3.1 Instructions for Testers and Their Attributes

In generating the ground-truth dataset, we requested three testers, aged between 25 and 35, who have more than five years of experience as fast-track career bureaucrats in the Japanese government, to perform the masking of the sample texts created in Step 1. All testers have experience in handling information disclosure requests, and one of them has extensive experience in processing a large number of such requests.

As shown in Step 2 of Figure 5.2, the testers were instructed to mask the characters containing confidential information within the sample texts by referring to Article 5 of the AAIHAO and the guidelines published by the Ministry of Internal Affairs and Communications of the Japanese government.

5.3.2 Masking Process Conducted by Testers

In the actual masking process, testers utilized an application designed to efficiently perform confidential information masking, as illustrated in Figure 5.3 and 5.4. Specifically, the testers accessed a web page implemented on Streamlit Cloud [39], where they processed 150 test text files. After opening a file in the web application, they selected the characters containing confidential information and pressed the corresponding confidentiality category button located at the bottom of the page. This action automatically masked the selected characters with a unique symbol assigned to each confidentiality category. Upon reviewing the entire text and completing the necessary masking, testers downloaded the masked text and submitted all 150 masked text files.

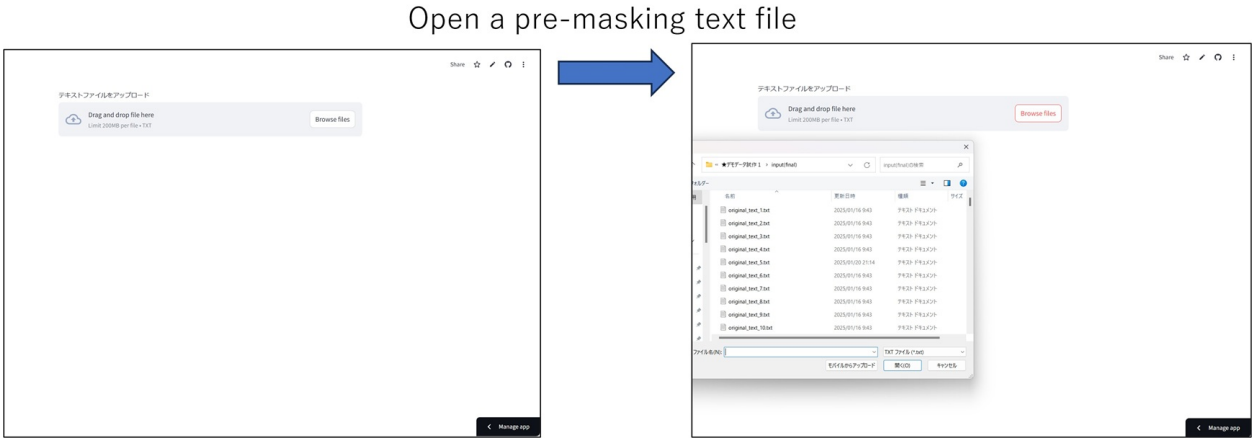

Figure 5.3: Application for masking task 1

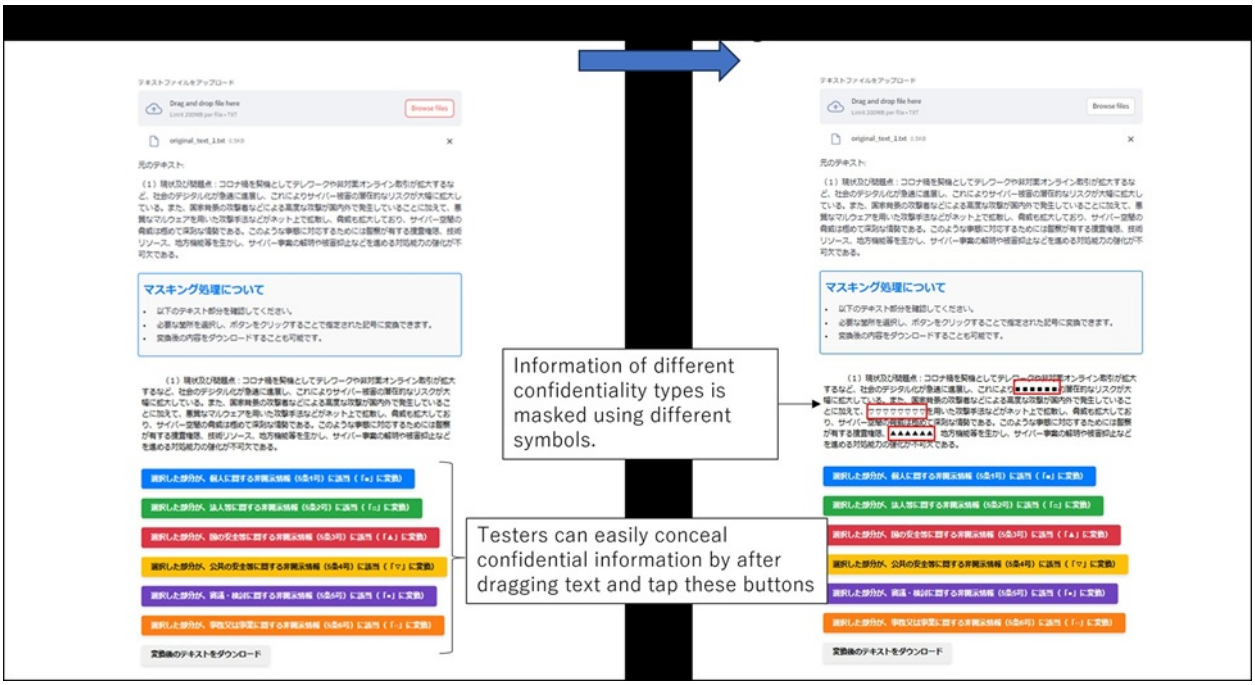

Figure 5.4: Application for masking task 2

5.3.3 Creation of Ground-Truth Tags from Review and Revisions

The testers' masking results were collected, and as shown in Step 3 of Figure 5.2, the creation of a single masking ground truth dataset was performed by integrating their making results. First, the masked text data submitted by each tester was reviewed by the authors. Any results that clearly deviated from the law or guidelines, or portions of the masking process that appeared to contain mistakes, were flagged for review, and the authors contacted the respective testers for clarification. However, any modifications to the masking results based on the review were left to the decision of the testers.

Subsequently, the masking results of the three testers were compared for each of the 150 text files. For any specific character in a particular file, if at least two testers out of the three performed the masking, that character was marked as masked in the ground truth dataset. Moreover, each tester explicitly tagged the confidentiality category associated with

each character they deemed confidential, using special symbols in the masking process in subsection 5.3.2. If the confidentiality category of a character to be masked was disputed by two or more testers, the category listed first in the AAIHAO Article 5 was chosen. This approach aligns with the general governmental criteria for determining whether information is confidential, based on the order in which the categories are listed in the law.

Thus, for all 150 text files, a list was generated indicating the correct masking positions and the confidentiality category (or 0 if the position was not deemed confidential) for each position in the text. An example of the created text data and its corresponding ground truth tags is shown in Figure 5.5.

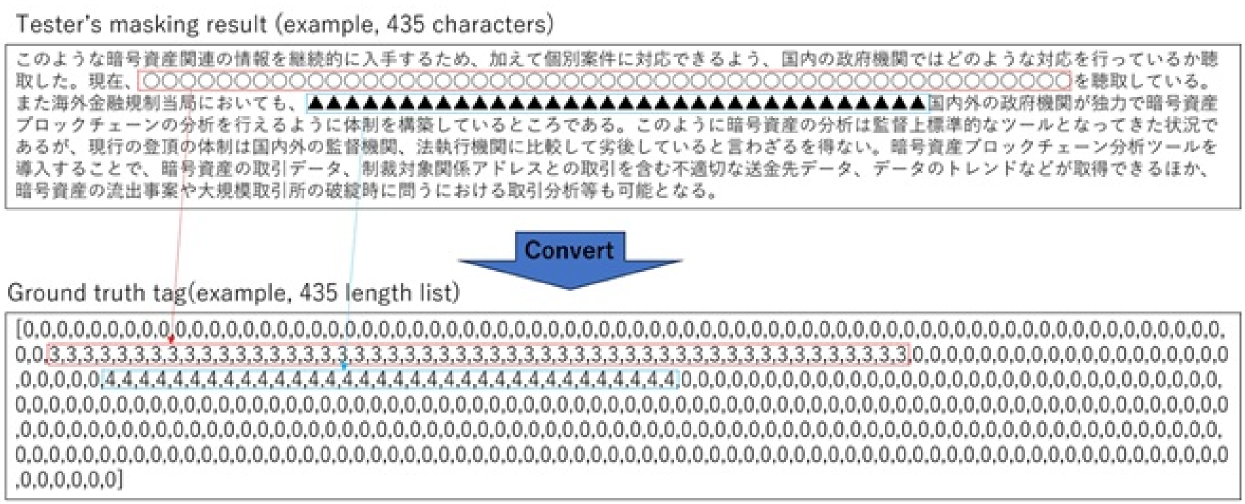

Figure 5.5: Example of ground truth tag

Correction Rate After Review for Each Tester

Table 5.2 presents the proportion of text files corrected by testers after reviewing the masking results of the first round for each text file. The review process incorporated feedback from the authors, particularly regarding apparent misapplications of the established criteria.

Notably, Tester B exhibited a significantly higher proportion of corrected files compared to

the other two testers. This was due to an observation that Tester B had over-interpreted the legal definition of corporate confidential information, leading to an overly broad masking scope, which was subsequently adjusted.

As a result, testers modified their decisions regarding whether each text file contained confidential information in approximately 12–21% of cases after review. Additionally, although the correction files for partial masking scope varied among testers, it exceeded 20% of all files. These findings indicate that even for human testers, the task of detecting and masking confidential information within this demonstration dataset was challenging and difficult to accomplish perfectly in a single iteration.

Tester	Modified Masking Scope Files	Modified Confidentiality Status Files
Tester A	51(34%)	29(19%)
Tester B	107(71%)	31(20%)
Tester C	34(23%)	19(13%)

Table 5.2: Correction details after review for each tester

5.4 Summary of the Created Test Dataset

The overview of the dataset, which consists of the ground truth data created by integrating the masking results of the three testers and the pre-masking text data, is shown in Figure 5.3, 5.4. Due to the workload involved in creating the test set, the masking text dataset consists of 150 sets. Furthermore, while referring to the actual proportions of confidential information summarized in Table 5.1, we ensured that the proportion of each confidentiality category did not fall below 5% of the total. Using this test data and ground truth dataset, we analyze the detecting and masking performance and trends of the LLM in Chapter 6.

In addition to the masked text data mentioned above, testers were also asked to score the subjective difficulty of masking each file on a scale from 1 to 5. Moreover, the number of

files modified after review process was recorded. These modifications included a variety of corrections, ranging from operational mistakes during masking to fundamental errors in masking judgment. Furthermore, we conducted interviews with each tester to estimate the approximate time required to mask the 150 text files. As shown in Table 5.5, manual masking by each tester took approximately 4.3 hours for 150 files, which corresponds to 1.72 minutes per file and about 0.2 seconds per character.

Table 5.3: Summary of Text Files and Samples

Total files	**Positive Sample**	**Negative Sample**	**Ave. Character**
150	103	47	421

Table 5.4: Confidential Category Breakdown in Positive Samples

Confidential Category	**Number of Cases**
1. Personal information (PII)	49 (47%)
2. Corporate information	30 (29%)
3. National security information	18 (17%)
4. Public safety information	17 (16%)
5. Deliberation/examination information	6 (6%)
6. Administrative operations information	12 (12%)

Table 5.5: Masking Time Required (Excluding Revisions)

Tester	**Time Required (hours)**
Tester A	6 hours
Tester B	3 hours
Tester C	4 hours
Average	**4.3 hours**

Chapter 6

Implementation and Evaluation

In this chapter, we evaluate the potential for improving the efficiency of government information disclosure requests using LLMs, utilizing the demo dataset created in Chapter 5. Specifically, we measure and assess the accuracy of two tasks related to information disclosure requests in government: detection and masking of confidential information within documents, using LLMs. Based on the results, we then analyze effective methods for utilizing LLMs to streamline information disclosure requests at this stage.

6.1 Overview of Implementation and Evaluation

Figure 6.1 shows an overview of the experimental patterns used in this study. In this research, we compare the results of tasks performed by LLMs with the ground truth dataset while varying 2–3 variables, and measure the accuracy of the detection and masking tasks.

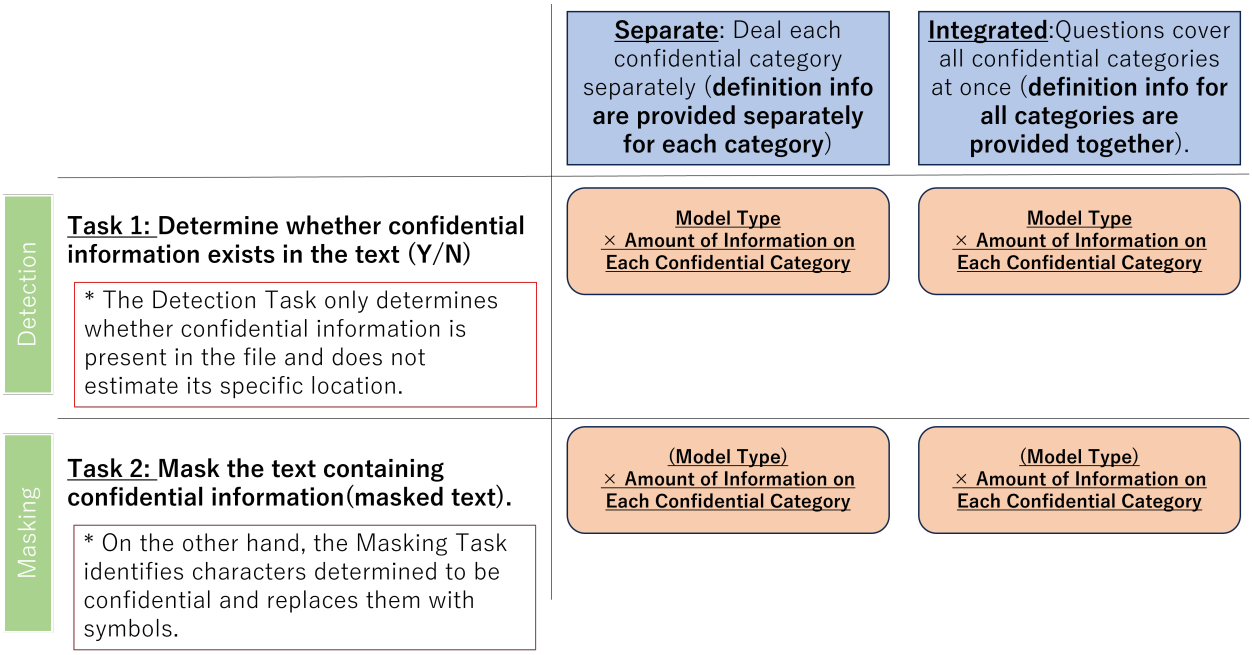

Figure 6.1: Experimental Patterns

Below, we provide detailed definitions of the detection and masking tasks, as well as a description of the three variables that were altered during the experiments.

6.1.1 Detailed Definitions of Detection and Masking Tasks

In this subsection, we define the detection and masking tasks. The expected outputs of the LLM for each task are shown in Table 6.1.

Table 6.1: Expected output of each task

Task	**Expected output**
Task 1. Detection	(Y/N):Estimation for presence of confidential information
Task 2. Masking	(Text):Text after masking for confidential information

Task 1: Detection Task

The detection task is the task of detecting whether specified confidential information exists within a given document file. The LLM is instructed to answer "Yes" or "No" as to whether the confidential information, defined within the prompt, is present in the text provided in the prompt. The key difference from the masking task is that in the detection task, the LLM is not required to identify which specific characters in the file correspond to confidential information. In other words, it only needs to estimate whether the file contains at least one character of confidential information, making this task simpler than the masking task. In practical workflows, if it is possible to quickly and accurately determine whether a document contains confidential information, this would contribute to the efficiency of confidential information screening in information disclosure requests. Therefore, this task was selected for the experiment. Additionally, while this study treats the detection and masking tasks independently, combining them in practice could potentially yield greater effectiveness.

Task 2: Masking Task

In the masking task, the LLM is given the text of a specific document and instructed to replace characters corresponding to specified confidential information with masking symbols. The key difference from the detection task is that the masking task requires the LLM to identify the exact characters in the document that constitute confidential information. As a result, the expected output of the LLM is the text after the masking has been applied. This makes the masking task more complex and challenging compared to the detection task.

Furthermore, the masking task closely resembles the manual redaction of confidential information currently performed in information disclosure requests. Even if the LLM can assist with part of the masking process or provide an initial draft, it could significantly improve the efficiency of actual document processing.

6.1.2 Confidential Information Categories

In this subsection, we explain the categories of confidential information contained in documents subject to detection and masking. In this study, with the aim of examining the potential efficiency improvements using LLMs in actual information disclosure requests within the Japanese government, we adopt the six types of confidential information exempted from disclosure under Article 5 of the AAIHAO, as presented in subsection 3.1.3, as the target confidential information for the detection and masking tasks. Table 6.2 summarizes these six types of confidential information again. For further details, please refer to subsection 3.1.3.

Table 6.2: Categories of Confidential Information

No.	Category
1	Personal information (PII)
2	Corporate information
3	National security information
4	Public safety information
5	Deliberation/examination information
6	Administrative operations information

6.1.3 Detailed Explanation about Variables in Tasks

In this subsection, we provide a detailed explanation of the variables considered when processing Task 1 and Task 2 with the LLM. Since this study serves as an initial investigation

into the applicability of LLMs for detecting and masking confidential information in government information disclosure, we prioritized simplicity. Accordingly, we conducted detection and masking experiments using a demo dataset while varying the following three variables.

1. Question Type: Ask integrated or separately

The following two strategies were tested for asking tasks. The differences in the way of asking with detailed prompts and the expected range of responses are summarized in Table 6.3.

1. **Separate**: A method to handle each category of confidential information individually.

2. **Integrated**: A method to handle all categories simultaneously.

These two approaches differ in how instructions are provided within the prompt, and whether the definition of confidential information is presented individually for each category or all at once. Based on the experimental results, we will analyze the differences in accuracy when handling confidential information individually versus simultaneously.

Table 6.3: Question type details

Question Type	Info. about confidential in prompt	Expected Answer Coverage
Separately	• Definitions/information of each confidential category (only the category being asked about is included)	• About each questioned category limited
Integrated	• Definitions for all confidential categories are included.	• About all confidential categories

2. Model Types

In this study, we mainly utilized the APIs provided by the Google Vertex AI platform (with OpenAI API also used in some experiments) to call LLMs and conduct experiments [15, 33]. The models used in these tasks will also be treated as variables for the experimental results. However, due to economic constraints related to API calls and time constraints related to token processing when handling large text files in various patterns, some experiments were limited to a smaller set of models. The models used and their settings are shown in Table 6.4.

Table 6.4: Models used in the Experiments

Model name	temperature	Platform
Claude-3-5-sonnet@20240620	0	Google virtex ai
Gemini-1.5-pro	0	Google virtex ai
Gemini-2.0-flash-exp	0	Google virtex ai
Llama-3.1-90b-vision-instruct-maas	0.1	Google virtex ai
Gpt-4o	0	OpenAI API

3. Amount of Information about Each Confidential Category

As the final variable, we examine how the amount of information provided within the prompt for each confidential information category affects performance. To maintain simplicity in our experimental design, we define two types of prompts, Prompt 1 and Prompt 2, which differ in the level of information provided regarding the definitions of confidential information. Specifically, we analyze how the amount of explanatory information included in the prompt influences the detection and masking of confidential information using a large language model (LLM). The details of the information provided in each prompt are as follows:

1. **Prompt 1 (Low-information case):** Only the legal definition from the law is provided.

2. **Prompt 2 (High-information case):** In addition to the legal definition from the law, a one-two line supplement describing the types of documents corresponding to each confidential category is included.

6.2 Performance Metrics

In this section, we describe the evaluation metrics used to assess the results of the confidential information detection and masking tasks performed by the LLMs on the demo dataset explained in Chapter 5. As shown in Figure 6.2, the evaluation metrics differ depending on the task type. The formulas for the metrics used are provided below.

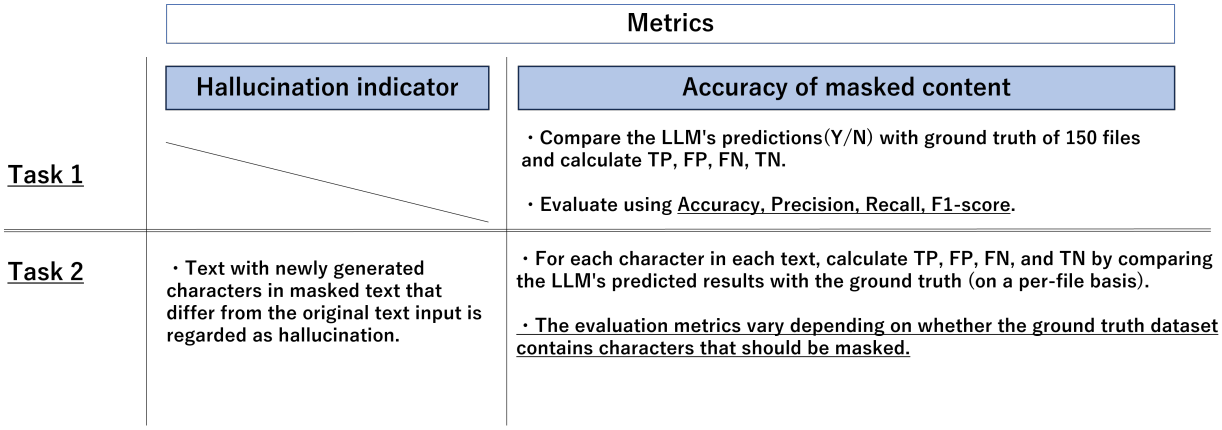

Figure 6.2: Performance Metrics Overview

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6.3}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.4}$$

$$\text{False Positive Rate} = \frac{FP}{FP + TN} \tag{6.5}$$

For the detection task, we refer to existing research and compare the LLM's predictions with the presence or absence of confidential information in the ground truth files. By comparing the predictions and ground truth tags for each file, we compute the values of True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). The evaluation is conducted using accuracy, precision, recall, and F1-score. The corresponding formulas for these metrics are shown above.

On the other hand, for the masking task, each character in a document file is tagged based on confidentiality. Based on these tags, we compute and sum the TP, FP, FN, and TN values for each document file. However, in some cases, there may be no characters requiring masking in the ground truth file. In such cases, TP and FN will necessarily be zero, causing precision and recall to be zero as well. This necessitates an adjustment in the evaluation metrics.

Considering this, the evaluation metrics differ based on whether a masking target file is a negative sample (no confidential information in ground truth) or a positive sample (contains confidential information in ground truth). As illustrated in Figure 6.3, the false positive rate is used to evaluate negative sample files, while precision and recall are used to evaluate

positive sample files.

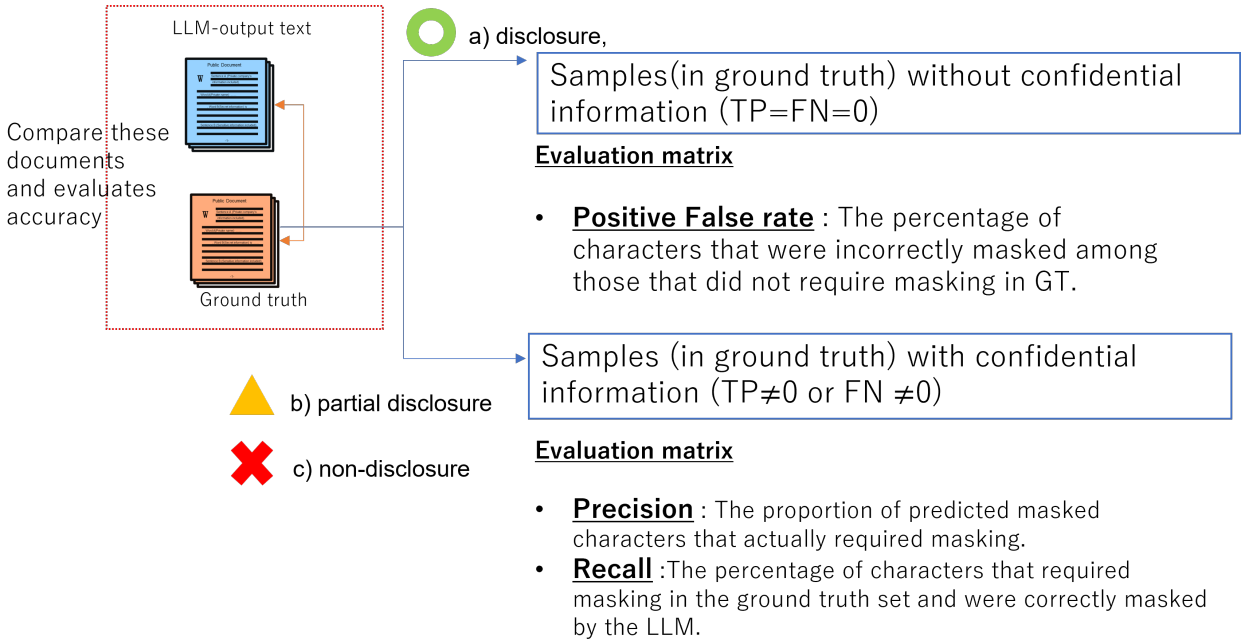

Figure 6.3: Performance Metrics Masking Task

6.3 Experimental Results: Task1 Detection

In this section, we summarize the results of the Detection task performed by the LLM. As mentioned earlier, in the Detection task, the LLM predicts whether the given text file contains any of the confidential categories listed in subsection 6.1.2. The accuracy of the predictions is evaluated by comparing the results with the ground truth.

6.3.1 Experimental Results of Individual Processing

First, in this section, we present the results of the detection task performed by the LLM for each confidential category. In the detection task for each category, the prompt provides

only information regarding the definition of the target confidential information, while no definitions or related information are given for non-target confidential information.

Figure 6.4 shows the detection results for each category of confidential information using Prompt1 for different models. Prompt1 represents a low-information case for each confidential category, containing only the legal definitions of each category.

Across the models, we can conclude that the detection performance for identifying the presence of Personally Identifiable Information (PII) is high. On the other hand, for categories 4 to 6, when only the legal definitions were provided, detection performance was insufficient. Particularly for categories 5 and 6, where recall is high but precision is low, the models tended to overestimate the presence of confidential information, labeling an excessive number of texts as containing confidential data.

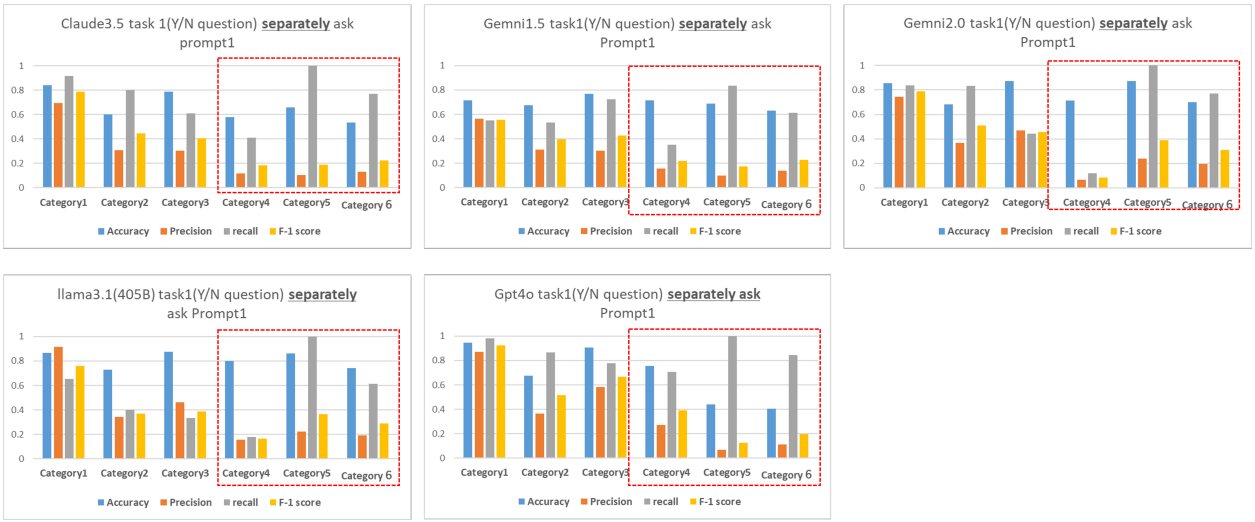

Figure 6.4: Task1 Separated Prompt1 Results

Next, we present the results of changes in detection performance when varying the Amount of Information about each confidential category under the same experimental settings. The results are shown in Figures 6.5, 6.6, 6.7, 6.8, and 6.9.

Comparing the accuracy of the detection task using Prompt1 (which contains only the legal

definition) and Prompt2 (which includes the legal definition along with an additional one to two lines of explanation), we observed that providing additional explanatory details about the confidential categories in Prompt2 improved overall detection accuracy. This improvement was particularly noticeable for confidential categories other than Category 1. Furthermore, a general trend of detection performance improvement was observed across all models tested in the experiment.

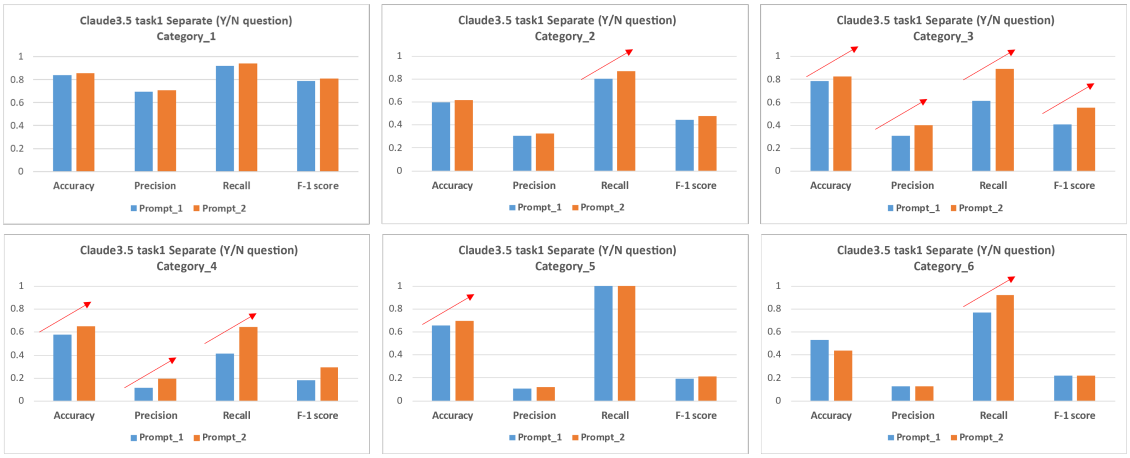

Figure 6.5: Task1 Separated Results based on Prompts (Clude3.5)

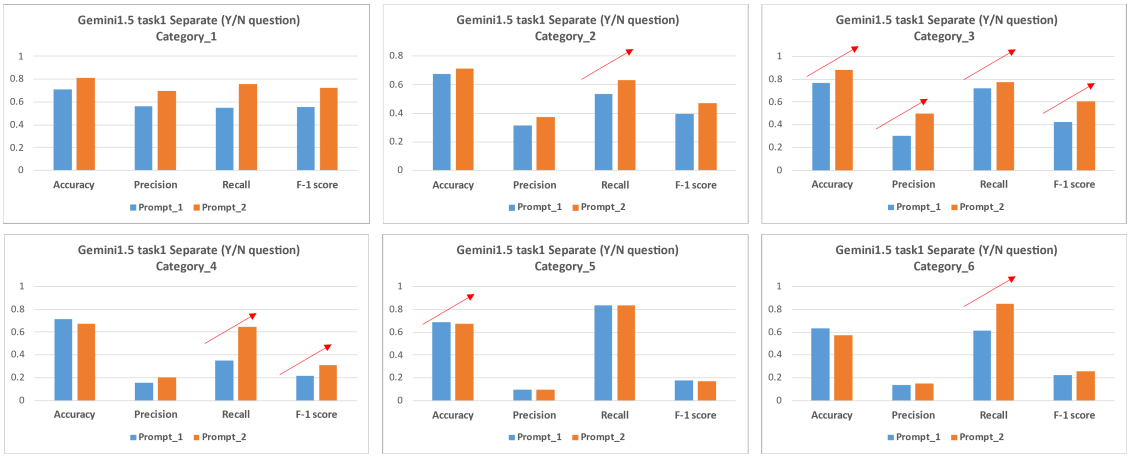

Figure 6.6: Task1 Separated Results based on Prompts (Gemini1.5)

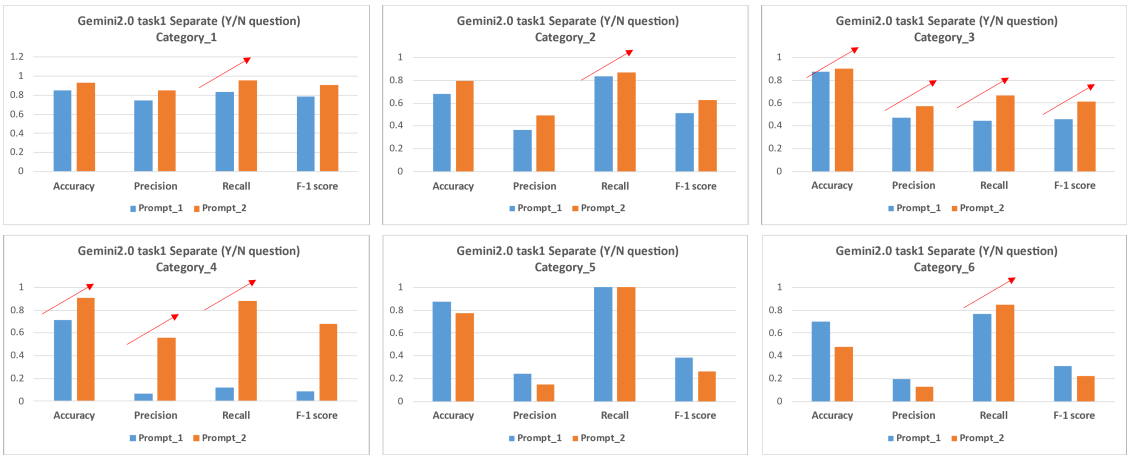

Figure 6.7: Task1 Separated Results based on Prompts (Gemini2.0)

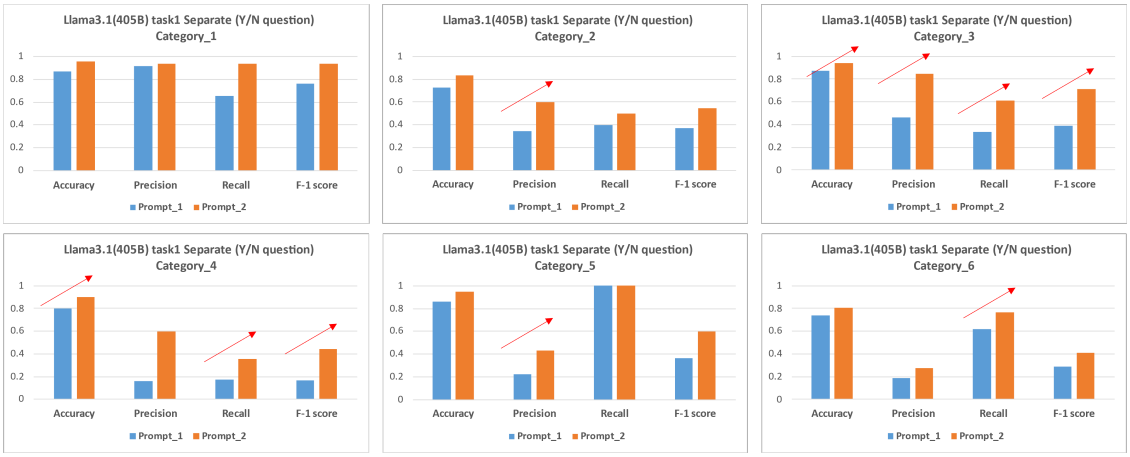

Figure 6.8: Task1 Separated Results based on Prompts (Llma3.1(405B))

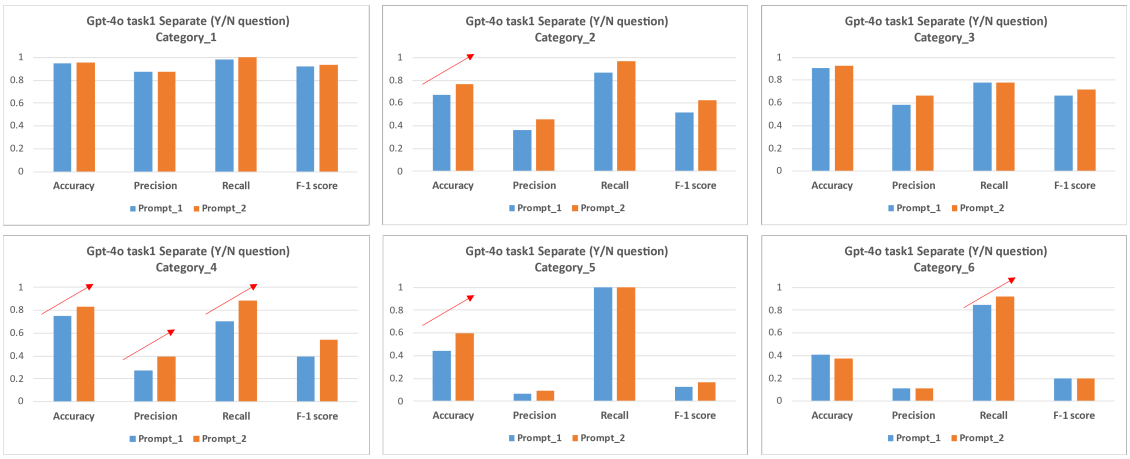

Figure 6.9: Task1 Separated Results based on Prompts (Gpt-4o)

6.3.2 Experimental Results of Simultaneous Processing

Next, this section presents the results when the detection task for the six confidentiality categories was performed simultaneously by the LLM. Unlike the task in Subsection 6.3.1, in this task, the model was instructed to output a simple ”yes” or ”no” response indicating whether confidential information was present, regardless of the specific confidentiality category. Additionally, the prompt provided definitions and information for all six confidentiality categories simultaneously.

Figure 6.10 compares the detection accuracy of each model when using Prompt 1 and Prompt 2. Furthermore, Figure 6.11 presents a comparison of detection accuracy for each model when using Prompt 1 and Prompt 2.

In this task, Claude and GPT-4 demonstrated very high accuracy in detecting files containing confidential information. With accuracy, recall, and precision all exceeding 0.8, utilizing LLMs for detecting confidential information in files is highly likely to be practical for real-world applications.

On the other hand, the open-source model Llama 3.1 (405B) exhibited lower detection performance compared to the other models. Moreover, as shown in Figure 6.11, increasing the amount of information provided about confidentiality categories in the prompt led to a slight improvement in detection performance in most models.

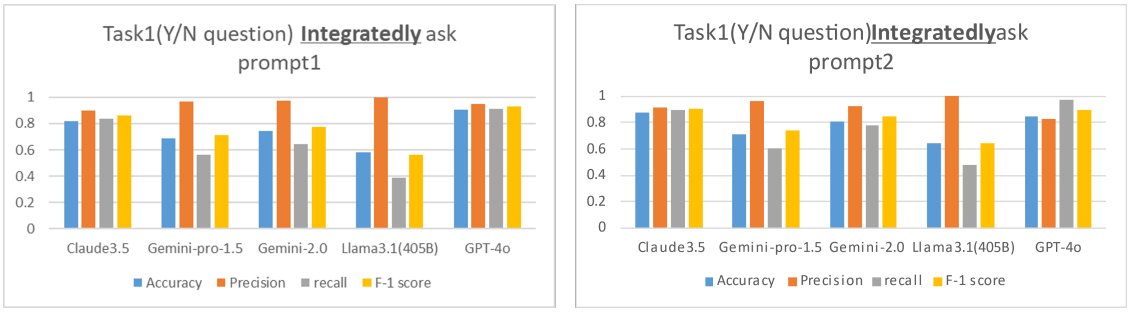

Figure 6.10: Task 1 Integrated Results based on Prompts (All models)

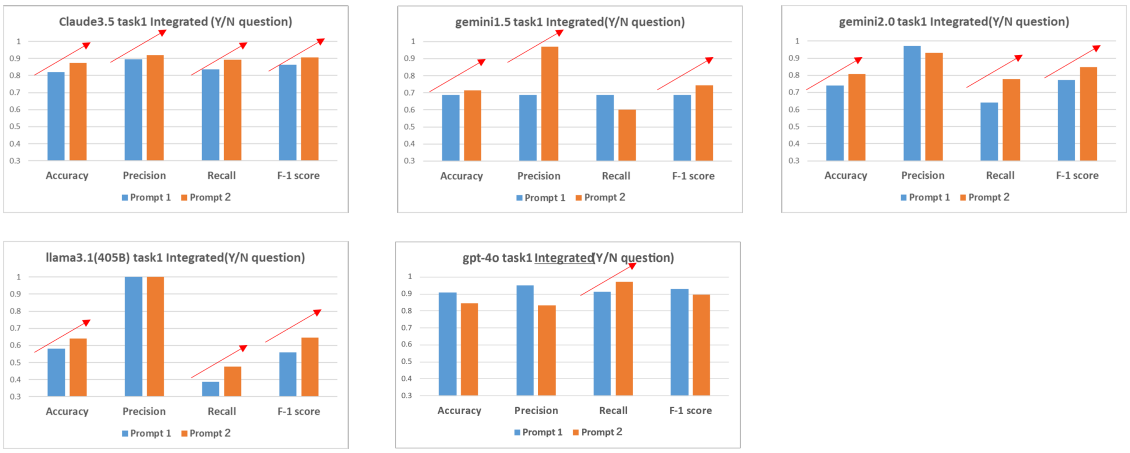

Figure 6.11: Task 1 Integrated Results based on Prompts (Each model)

6.4 Experimental Results: Task2 Masking

This section summarizes the results of the masking task performed by the LLM. As mentioned earlier, in the masking task, the LLM is instructed to replace specific characters in a given text file with designated masking symbols if the text contains any of the confidentiality categories listed in Subsection 6.1.2. Finally, the accuracy of the LLM's masking decisions is evaluated by comparing the masked characters in the LLM's output with the ground truth dataset.

6.4.1 Post Processing for LLM Output Masked Text

This section explains the post-processing applied to the text output by the LLM, where confidential information has been replaced with specific masking characters.

One characteristic of LLMs when masking Japanese documents is that the number of masked characters before masking sometimes does not match the number of characters corresponding to the masking symbols. As shown in Figure 6.12, this masking behavior is appropriate, except for the discrepancy in the number of masking symbols in each masked part. However,

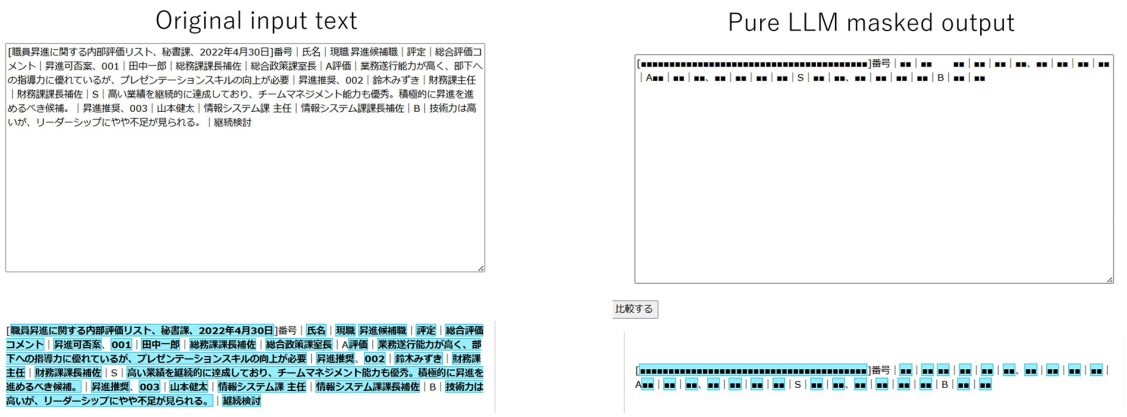

Figure 6.12: Post-processing for LLM output masked text 1

as mentioned in Subsection 5.3.3, since the ground truth dataset tracks the exact positions of the characters that need to be masked using position indexes, a process was required to restore the character count changes in the masked portions back to the original string to compare the LLM output with the ground truth data. This adjustment is likely a specific consideration unique to experiments using Japanese text.

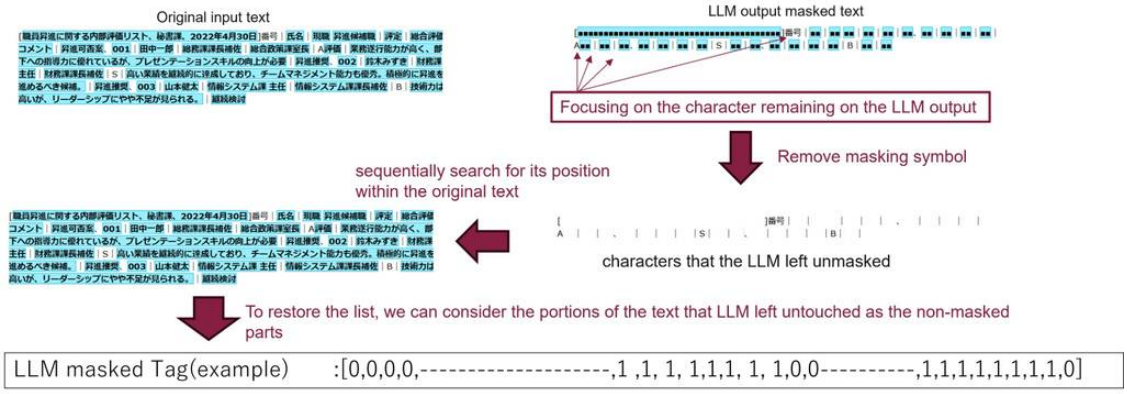

Figure 6.13: Post-processing for LLM output masked text 2

Process 1: First, we check whether the masked output by the LLM contains any characters that were not present in the original text. If such characters are found, we consider it a case of hallucination and exclude it from the accuracy evaluation.

Process 2: Next, for each character that remains unmasked in the LLM's output string, we

perform a sequence search to determine its position in the original text. This allows us to identify which characters in the LLM's masked output correspond to the indexed positions in the original text.

Process 3: Finally, we generate a list of the same length as the original text, assigning a value of 0 to the indices where characters remain and 1 to all other positions. This list retains the LLM's masking results in a format that corresponds to the index positions of the original input string.

We then compare this list with the ground truth list to evaluate the accuracy of the LLM's masking. While most of the process is automated, manual verification was performed when necessary to ensure the correctness of the processing.

6.4.2 Experimental Results of Individual Processing

This section presents the results of the masking task performed by the LLM for each confidentiality category, separately. For the masking task, we compared the LLM's output for each file (after applying the post-processing described in Section 6.4.1) with the ground truth, calculated TP, FN, FP, and TN, and used these values to compute evaluation metrics for each file.

As described in Section 6.2, positive samples were evaluated using precision and recall, while negative samples were evaluated using the false positive rate. Additionally, files identified as containing hallucinations during post-processing were excluded from the evaluation calculations, and their count was recorded separately.

Figures 6.14, 6.15, 6.16, and 6.17 show the prediction evaluation results for each category across different models. In each figure, the left side represents the evaluation results using Prompt 1, while the right side shows the results using Prompt 2. The top section displays

the evaluation results for positive samples, followed by the evaluation results for negative samples, and the bottom section presents the number of files identified as hallucinations during the evaluation process.

When comparing the models, Claude 3.5 demonstrated overall higher masking accuracy than the other models. Additionally, the number of hallucinations generated by Claude 3.5 was lower compared to the other models.

For Gemini 1.5 and LLaMA 3.1 (405B), while their precision for positive samples was relatively high, their recall was notably lower. This suggests that these models tend to overestimate the presence of confidential information, resulting in excessive masking of characters that may not actually contain confidential content.

Focusing on the impact of using Prompt 1 versus Prompt 2, we observed a trend in Claude 3.5 and Gemini 2.0 where precision slightly improved across multiple confidential categories, while recall increased significantly. In the case of Claude 3.5, the best-performing model, the use of Prompt 2 resulted in precision exceeding 0.6 and recall reaching approximately 0.5 for many categories.

Regarding the false positive rate for negative samples, switching to Prompt 2 did not lead to consistent improvements; in some cases, it improved, while in others, it worsened. Similarly, the number of hallucinations did not show a significant difference between Prompt 1 and Prompt 2.

Figure 6.14: Task2 Separated Results based on Prompts (Claude3.5)

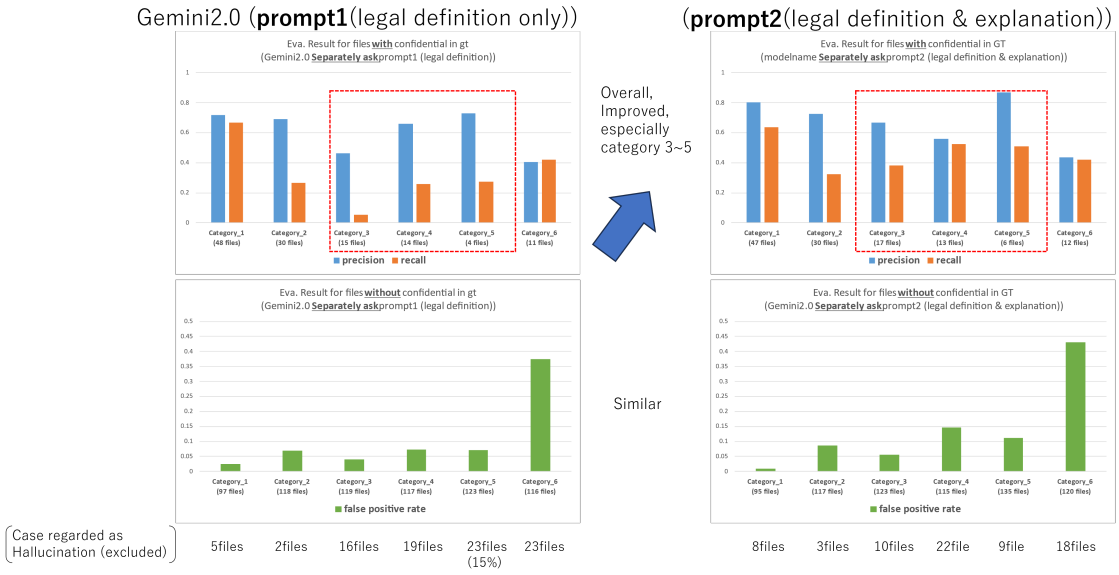

Figure 6.15: Task2 Separated Results based on Prompts (Gemini2.0)

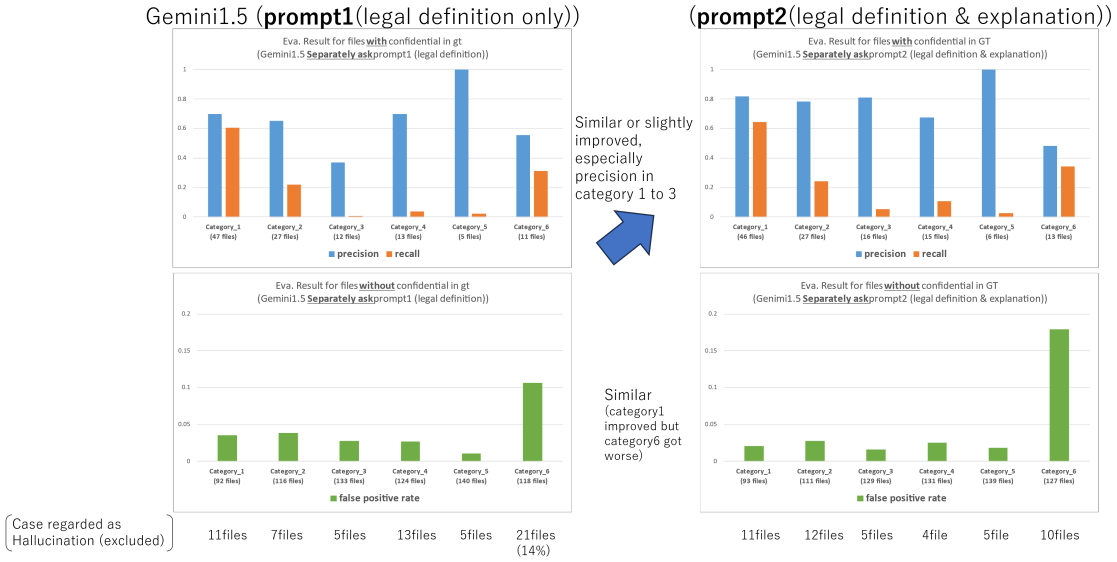

Figure 6.16: Task2 Separated Results based on Prompts (Gemini1.5)

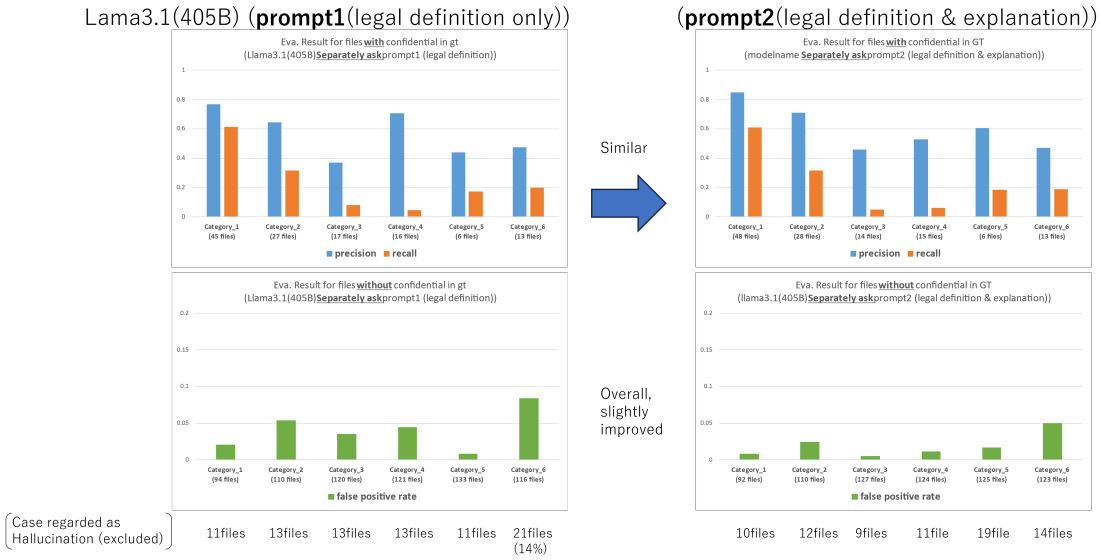

Figure 6.17: Task2 Separated Results based on Prompts (Llama3.1(405B))

6.4.3 Experimental Results of Simultaneous Processing

The results of the masking task when applied collectively to all confidential categories are shown in Figure 6.18.

We compared the performance of Prompt 1 and Prompt 2 when providing definitions for all sensitive categories and instructing the LLMs to mask all relevant sensitive information. For Claude 3.5 and LLaMA 3.1 (405B), there were no significant changes in precision or recall for positive samples. However, the false positive rate for negative samples slightly worsened in Claude 3.5.

On the other hand, for Gemini 2.0 and Gemini 1.5, precision for positive samples remained almost unchanged, while recall showed a slight improvement. However, this was accompanied by a deterioration in the false positive rate for negative samples. Thus, simply increasing the amount of information in the prompt did not necessarily lead to improved masking accuracy.

Across all models, precision was around 0.6, while recall ranged between 0.25 and 0.4. Furthermore, the number of files identified as hallucinations increased compared to cases where masking was performed separately for each category.

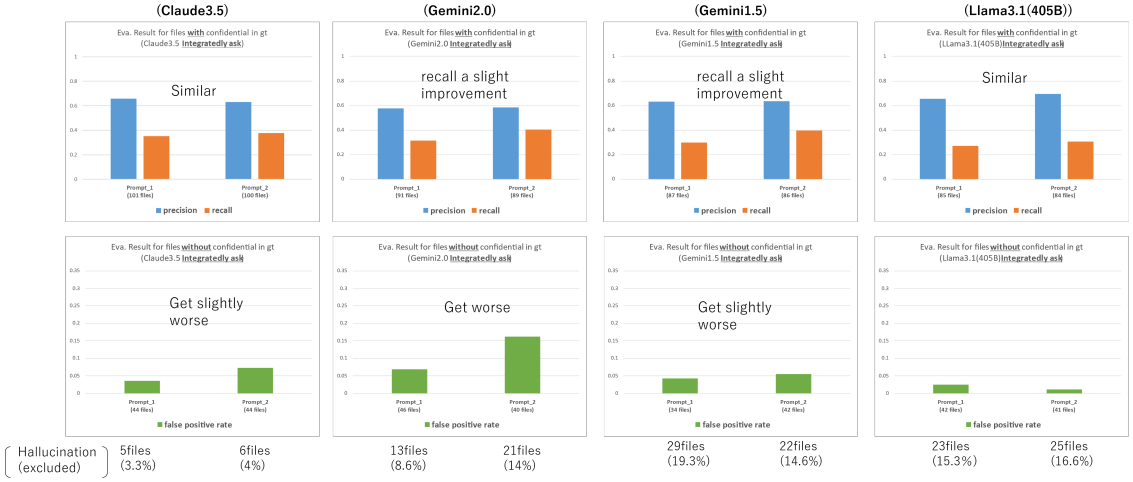

Figure 6.18: Task2 Integrated Results based on Prompts

6.5 Discussions

In this chapter, we evaluate the performance of detecting and masking confidential information using LLMs through experiments with an original Japanese text dataset. Specifically, we adopt the Japanese government's information disclosure process as a model case and instruct LLMs to perform detection and masking tasks for six categories of confidential information as defined by related regulations. The results are then compared with a ground truth dataset created by human testers for evaluation. In our experiments, we conduct detection and masking tasks under two conditions: processing each of the six sensitive categories separately and processing them simultaneously. We investigate how the type of model and the amount of information provided in the prompt regarding confidential categories affect the accuracy of detection and masking.

The experimental results indicate that when processing each category separately, the accuracy varied across categories. Notably, the detection and masking accuracy for Personally Identifiable Information (PII), category 1, remained high even when only the legal definitions were provided in the prompt. In contrast, the accuracy for information related to corporations' secret (Category 2), national security (Category 3), public safety (Category 4), deliberative and consultative information (Category 5), and public administration and business information (Category 6) was relatively low. However, when increasing the amount of information provided in the prompt regarding confidential information, we observed an improvement in accuracy for both detection and masking tasks, particularly with models such as Claude 3.5 and Gemini 2.0. These results suggest that for confidential information other than PII, providing only legal definitions is insufficient for accurate identification by LLMs. Instead, additional information or specific examples in the prompt are necessary to provide LLMs with sufficient context.

For the detection task, even when all confidential information categories were processed simultaneously, increasing the amount of information in the prompt led to improved accuracy across all models. Claude 3.5 and GPT-4o achieved accuracy, precision, and recall values exceeding 0.8, successfully detecting files containing confidential information with high accuracy. These models outperformed others such as Gemini 1.5, Gemini 2.0, and LLaMA 3.1 (405B). These findings suggest that selecting high-performance models and providing sufficient definitions of confidential information in the prompt can effectively enable LLMs to detect the presence of confidential information in text files.

On the other hand, in the masking task, we observed that LLMs occasionally output text containing characters different from the original text, excluding the masking characters. Detecting and correcting such hallucinations is crucial when using LLMs in the confidentail information processing stage of an information disclosure in the governments. Our experiments revealed significant differences in hallucination occurrence rates among models, with a more than twofold difference observed between Claude 3.5 and Gemini 2.0. While a detailed analysis of hallucination patterns remains a future research topic, linguistic characteristics of the training data may have contributed to the higher hallucination rates in certain models during our Japanese text-based experiments.

From a model comparison perspective, the masking task exhibited substantial differences in performance when processing confidential categories separately. Overall, Claude 3.5 demonstrated superior performance compared to other models due to its low hallucination rate and high masking accuracy. Conversely, models such as Gemini 1.5 and LLaMA 3.1 (405B) did not achieve sufficient accuracy in masking confidential information other than PII (Category 1). Furthermore, apart from LLaMA 3.1 (405B), all other models exhibited a higher false positive rate when masking public administration and business information (Category 6) compared to other categories. This may be due to overly broad definitions of confidential

information provided in the prompt, leading to an increased false positive rate. However, the reason why LLaMA 3.1 (405B) exhibited a lower false positive rate remains an open question for future research.

When all confidential categories were masked simultaneously, the precision across all models was around 0.6, while recall ranged between 0.3 and 0.4. However, compared to masking each category separately, most models showed limited improvement in masking accuracy as the amount of confidential information category in the prompt increased, or in some cases, the false positive rate worsened. We speculate that masking specific confidential characters in text is inherently a complex task, and increasing the amount of information regarding confidential categories further added to this complexity, thereby hindering the performance improvement of LLMs.

For the masking task, using Claude 3.5 to process each confidential category separately (using Prompt 2) resulted in higher masking accuracy across all models and compared to masking all categories at once. This suggests that instructing LLMs to mask each confidential category individually and increasing the amount of definitional information in the prompt is an effective strategy for the masking task. However, it is also necessary to consider the false positive rate in negative samples when implementing this approach.

Finally, regarding the applicability of LLMs to masking confidential information in response to information disclosure requests, even under the best-performing setting—where each confidential category was masked individually by using claude3.5 and prompt2—the results showed that, except for Category 1 (PII), precision remained around 0.6–0.7, while recall was approximately 0.4–0.6. These findings suggest that, given the models and prompts used in this study, employing LLMs for masking tasks is best suited as a support tool for human reviewers rather than as a fully automated solution. In particular, since recall was only around 0.4–0.6, this indicates that nearly half of the information that should have

been masked was overlooked. Therefore, it is essential to establish a verification process in which human reviewers check and refine the LLM-generated masking results to ensure accuracy. Given the substantial volume of documents that need to be processed in governmental information disclosure requests, leveraging LLMs as a first-stage screening tool, followed by human verification, appears to be a highly promising approach to improving overall operational efficiency.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis, we investigate the potential of using LLMs to detect and mask confidential information in text, aiming to streamline the response process to information disclosure requests in the government. The study specifically evaluates the use of LLMs for the tasks of detecting and masking confidential information by using the Japanese government's information disclosure process as a case study. A custom Japanese-language dataset was created, and a ground truth dataset was developed with the assistance of testers who had experience working in government. The dataset was then used to evaluate LLM performance in detection and masking tasks for six categories of confidential information defined in the regulations. The approach involved varying models, instructions, and the amount of information regarding confidential categories in the prompts to assess the accuracy by comparing the LLM outputs with human-created ground truth.

In the detection task, the highest accuracy pattern achieved over 80% accuracy, precision, and recall in correctly identifying files containing confidential information. Additionally, increasing the amount of information regarding confidential information in the prompt improved accuracy, indicating that prompt engineering could enhance detection performance. This demonstrates the high potential of applying LLMs to the detecting whether or not confidential information exist in text files, suggesting that adoption of LLMs in detection

task in the public sector should be actively pursued.

In the masking task, the accuracy was lower than in the detection task, revealing that this task is more complex and challenging for LLMs. Moreover, the improvement in accuracy by increasing the information about confidential categories in the prompt was smaller for the masking task than for the detection task. However, the best-performing model achieved precision of around 0.7–0.8 and recall of approximately 0.4–0.6 for most confidential information categories when masking each category individually. This suggests that with careful prompt design, LLMs could serve as a valuable tool for drafting confidential information masking or supporting workers in the process. It is also worth noting that in the masking task, hallucinations could occur in the LLM-generated text, which highlights the importance of detecting and correcting these hallucinations.

As mentioned above, the results indicate a promising direction for streamlining government information disclosure processes with LLM-based approaches for confidential information management.

7.2 Future Work

Although this results of this study shows a lot of findings as initial study, several areas for expansion remain, which serve as future research challenges.

7.2.1 Experiments with a Broader Range of Models

Due to financial constraints, our experiments are conducted using the Google Vertex AI platform and certain OpenAI APIs, which limited the models available for testing [15, 33]. Recently, more advanced models have been released, and utilizing the latest models may

enable more accurate detection and masking of confidential information. Additionally, conducting experiments with a broader range of LLMs would allow for a more detailed analysis of performance differences across models.

7.2.2 Experiments with Larger and More Realistic Datasets

In this study, we created a custom dataset replicating Japanese government administrative documents for our experiments. When applying LLMs to detect and mask confidential information in government information disclosure requests, it is preferable to conduct performance evaluations using larger datasets that closely resemble actual operational data. Future research should focus on using datasets that more accurately reflect real government documents and conducting experiments with a large volume of data to obtain more detailed results.

7.2.3 Enhancing Workflow Efficiency in Human-LLM Collaboration

The results of this experiment indicate that for the masking task, LLMs can be effectively utilized as a support tool for human reviewers rather than as a standalone solution. When considering the practical implementation of LLMs in governmental information disclosure processes, it is essential to carefully design a workflow that integrates both LLM-assisted processing and human verification. At present, a feasible approach would be to use LLMs for an initial screening of a large volume of documents requiring redaction, followed by human review to ensure accuracy. Additionally, further automation could be achieved by incorporating detection tasks alongside masking or by having an independent LLM verify the redaction draft generated by another LLM.

However, the risk of incorrect judgments or hallucinations in LLM-generated outputs remains a concern. Therefore, it is crucial to establish a robust detection process for such errors. Ultimately, the responsibility for finalizing redacted documents should rest with human officers who oversee document management, ensuring that the final outputs meet the required standards of accuracy and compliance.

7.2.4 Potential Model Optimization through Fine-Tuning

Finally, while this study treated the method of providing information in prompts and task instructions as variables, we utilized basic models provided via API. Given the specialized task of processing confidential information in government information disclosure, fine-tuning models using dedicated datasets may contribute to improved detection and masking accuracy. Exploring the efficiency gains of using fine-tuned models tailored for confidential information processing in government disclosure remains a future research challenge. In particular, given the vast amount of government documents and past information disclosure request responses, considering the option of fine-tuning models is a crucial factor when evaluating the potential use of LLMs in these operations.

Bibliography

[1] Ai4Privacy. Pii masking 300k, 2024. URL https://huggingface.co/datasets/ai4privacy/pii-masking-300k. Accessed: 2025-01-30.

[2] Federico Albanese, Daniel Ciolek, and Nicolas D'Ippolito. Text sanitization beyond specific domains: Zero-shot redaction & substitution with large language models, 2023. URL https://arxiv.org/abs/2311.10785.

[3] Jason R. Baron, Mahmoud F. Sayed, and Douglas W. Oard. Providing more efficient access to government records: A use case involving application of machine learning to improve foia review for the deliberative process privilege. *J. Comput. Cult. Herit.*, 15 (1), January 2022. ISSN 1556-4673. doi: 10.1145/3481045. URL https://doi.org/10.1145/3481045.

[4] Jason R Baron, Nathaniel W Rollings, and Douglas W Oard. Using chatgpt for the foia exemption 5 deliberative process privilege. In *LegalAIIA@ ICAIL*, pages 32–48, 2023.

[5] K. Branting, B. Brown, C. Giannella, et al. Decision support for detecting sensitive text in government records. *Artificial Intelligence and Law*, 33(2):171–197, 2025. doi: 10.1007/s10506-023-09383-6. URL https://doi-org.ezproxy.lib.vt.edu/10.1007/s10506-023-09383-6.

[6] Jonathan Bright, Florence Enock, Saba Esnaashari, John Francis, Youmna Hashem, and Deborah Morgan. Generative ai is already widespread in the public sector: evidence from a survey of uk public sector professionals. *Digit. Gov.: Res. Pract.*, October 2024. doi: 10.1145/3700140. URL https://doi.org/10.1145/3700140. Just Accepted.

[7] Chun Jie Chong, Chenxi Hou, Zhihao Yao, and Seyed Mohammadjavad Seyed Talebi. Casper: Prompt sanitization for protecting user privacy in web-based large language models, 2024. URL https://arxiv.org/abs/2408.07004.

[8] McKinsey & Company. The economic potential of generative ai: The next productivity frontier, 2023. URL https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction. Accessed: 2025-01-09.

[9] Department of Homeland Security. Department of homeland security unveils artificial intelligence roadmap, announces pilot projects to maximize benefits of technology, advance homeland security mission. https://www.dhs.gov/archive/news/2024/03/18/department-homeland-security-unveils-artificial-intelligence-roadmap-announces, 2024. Accessed: 2025-01-21.

[10] Digital Agency, Government of Japan. In fiscal 2023, we conducted a technology validation for the appropriate use of generative ai in digital agency and administration, 2024. URL https://www.digital.go.jp/en/news/19c125e9-35c5-48ba-a63f-f817bce95715. Accessed: 2025-01-21.

[11] European Union. AI Act: Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Published in the Official Journal of the European Union, 2024. URL https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[12] Keyuan Fang and Kewei Xu. Automating government response to citizens'questions: A large language model-based question-answering guidance generation system. In *2023*

3rd International Conference on Digital Society and Intelligent Systems (DSInS), pages 386–389, 2023. doi: 10.1109/DSInS60115.2023.10455136.

[13] G7. G7 Leaders' Statement on the Hiroshima AI Process. Online; accessed January 20, 2025, October 2023. URL https://www.mofa.go.jp/ecm/ec/page5e_000076.html. Issued on October 30, 2023. Available in Japanese and English.

[14] Aitor García Pablos, Naiara Perez, and Montse Cuadros. Sensitive data detection and classification in Spanish clinical text: Experiments with BERT. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4486–4494, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.552/.

[15] Google Cloud. Vertex ai, 2025. URL https://cloud.google.com/vertex-ai?hl=en. Accessed: 2025-02-13.

[16] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. URL https://arxiv.org/abs/2312.06674.

[17] Information Commissioner's Office. Guide to managing an foi request, 2025. URL https://ico.org.uk/for-organisations/foi/guide-to-managing-an-foi-request/. Accessed: 2025-01-23.

[18] Information Commissioner's Office. Request handling, freedom of information – frequently asked questions. urlhttps://ico.org.uk/for-organisations/foi/freedom-of-information-and-environmental-information-regulations/section-45-code-of-practice-request-handling/request-handling-freedom-of-information-frequently-asked-questions/, 2025. Accessed: 2023-01-23.

[19] Japanese Government. Act on access to information held by administrative organs, 1999. URL https://www.japaneselawtranslation.go.jp/ja/laws/view/3765. Accessed: 2023-01-23.

[20] The Learning Agency Lab. Pii data detection, 2024. URL https://www.kaggle.com/competitions/pii-detection-removal-from-educational-data/overview. Accessed: 2025-01-30.

[21] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3197–3207, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539147. URL https://doi.org/10.1145/3534678.3539147.

[22] Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. Deid-gpt: Zero-shot medical text de-identification by gpt-4, 2023. URL https://arxiv.org/abs/2303.11032.

[23] Jörn Von Lucke and Sander Frank. A few thoughts on the use of chatgpt, gpt 3.5, gpt-4 and llms in parliaments: Reflecting on the results of experimenting with llms in the parliamentarian context. *Digit. Gov.: Res. Pract.*, May 2024. doi: 10.1145/3665333. URL https://doi.org/10.1145/3665333. Just Accepted.

[24] Microsoft. Azure ai content safety. https://learn.microsoft.com/en-us/azure/ai-services/content-safety/, 2024. Accessed: 2025-01-22.

[25] Ministry of Internal Affairs and Communications, Government of Japan. Information

disclosure system implementation status survey, 2025. URL https://www.soumu.go.jp/main_sosiki/gyoukan/kanri/jyohokokai/chousa.html. Accessed: 2025-01-26.

[26] Ministry of Internal Affairs and Communications, Japan. Information Disclosure and Personal Information Protection: Advisory and Judicial Precedent Database, 2025. URL https://koukai-hogo-db.soumu.go.jp/. Accessed: 2025-01-30.

[27] Sho Mori and Atsuya Kato. Exploration of automated identification of trade secret information using large language models. In *Proceedings of the Computer Security Symposium 2024*, pages 805–812, Kobe, Japan, oct 2024. Information Processing Society of Japan. URL http://id.nii.ac.jp/1001/00240735/.

[28] Gabriel Nicholas and Paul Friedl. Regulating large language models: A roundtable report, 2024. URL https://arxiv.org/abs/2403.15397.

[29] NVIDIA Corporation. What are large language models?, 2025. URL https://www.nvidia.com/en-us/glossary/large-language-models/. Accessed: 2025-03-05.

[30] Government of U.K. Freedom of information act 2000. https://www.legislation.gov.uk/ukpga/2000/36/contents, 2000. Accessed: 2025-01-23.

[31] Government of U.K. How to make a freedom of information (foi) request. https://www.gov.uk/make-a-freedom-of-information-request, 2025. Accessed: 2025-01-23.

[32] OpenAI. Openai moderation. https://platform.openai.com/docs/guides/moderation, 2024. Accessed: 2025-01-22.

[33] OpenAI. Openai api, 2025. URL https://openai.com/index/openai-api/. Accessed: 2025-02-13.

[34] Alejo Paullier. Pii | external dataset, 2023. URL https://www.kaggle.com/datasets/alejopaullier/pii-external-dataset. Accessed: 2025-01-30.

[35] Sanjeev Pulapaka, Srinath Godavarthi, and Sherry Ding. *Empowering the Public Sector with Generative AI: From Strategy and Design to Real-World Applications*. Apress, 2024. ISBN 9798868804731. Released July 2024, Available on the O'Reilly learning platform with a 10-day free trial.

[36] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Large language models are advanced anonymizers, 2025. URL https://arxiv.org/abs/2402.13846.

[37] Dimitri Staufer, Frank Pallas, and Bettina Berendt. Silencing the risk, not the whistle: A semi-automated text sanitization tool for mitigating the risk of whistleblower re-identification. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 733–745, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658936. URL https://doi.org/10.1145/3630106.3658936.

[38] Vincent J. Straub, Youmna Hashem, Jonathan Bright, Satyam Bhagwanani, Deborah Morgan, John Francis, Saba Esnaashari, and Helen Margetts. Ai for bureaucratic productivity: Measuring the potential of ai to help automate 143 million uk government transactions, 2024. URL https://arxiv.org/abs/2403.14712.

[39] Streamlit. Streamlit community cloud. https://streamlit.io/cloud, 2025. Accessed: 2025-01-30.

[40] Xiongtao Sun, Gan Liu, Zhipeng He, Hui Li, and Xiaoguang Li. Deprompt: Desensitization and evaluation of personal identifiable information in large language model prompts, 2024. URL https://arxiv.org/abs/2408.08930.

[41] U.K. Department for Science, Innovation and Technology. Government's experimental ai chatbot to help people set up small businesses and find support. https://www.gov.uk/government/news/governments-experimental-ai-chatbot-to-help-people-set-up-small-businesses-and-find-support, 2024. Accessed: 2025-01-21.

[42] U.S. Congress. Freedom of information act, 1966. URL https://www.foia.gov/. Accessed: 2025-01-23.

[43] U.S. Department of Homeland Security. Artificial intelligence use case inventory, 2024. URL https://www.dhs.gov/ai/use-case-inventory. Accessed: 2025-01-21.

[44] U.S. Department of Justice. Foia.gov - freedom of information act: Create an annual report. https://www.foia.gov/data.html, 2024. Accessed: 2025-01-23.

[45] Peter Viechnicki and William D. Eggers. How much time and money can ai save government?, 2017. URL https://www2.deloitte.com/content/dam/insights/us/articles/3834_How-much-time-and-money-can-AI-save-government/DUP_How-much-time-and-money-can-AI-save-government.pdf.

[46] Shenao Wang, Yanjie Zhao, Xinyi Hou, and Haoyu Wang. Large language model supply chain: A research agenda. *ACM Trans. Softw. Eng. Methodol.*, December 2024. ISSN 1049-331X. doi: 10.1145/3708531. URL https://doi.org/10.1145/3708531. Just Accepted.

[47] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023. doi: 10.1109/JAS.2023.123618.

[48] Jianliang Yang, Xiya Zhang, Kai Liang, and Yuenan Liu. Exploring the application of

large language models in detecting and protecting personally identifiable information in archival data: A comprehensive study*. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2116–2123, 2023. doi: 10.1109/BigData59044.2023.10386949.