Privacy Meets Explainability: Managing Confidential Data and Transparency Policies in LLM-Empowered Science

Yashothara Shanmugarasa[*]
yashothara.shanmugarasa@data61.csiro.au
CSIRO's Data61
Sydney, NSW, Australia

Shidong Pan[†]
Shidong.Pan@data61.csiro.au
CSIRO's Data61
Sydney, NSW, Australia

Ming Ding
ming.ding@data61.csiro.au
CSIRO's Data61
Sydney, NSW, Australia

Dehai Zhao
dehai.zhao@data61.csiro.au
CSIRO's Data61
Sydney, NSW, Australia

Thierry Rakotoarivelo
thierry.rakotoarivelo@data61.csiro.au
CSIRO's Data61
Sydney, NSW, Australia

Abstract

As Large Language Models (LLMs) become integral to scientific workflows, concerns over the confidentiality and ethical handling of confidential data have emerged. This paper explores data exposure risks through LLM-powered scientific tools, which can inadvertently leak confidential information, including intellectual property and proprietary data, from scientists' perspectives. We propose "DataShield", a framework designed to detect confidential data leaks, summarize privacy policies, and visualize data flow, ensuring alignment with organizational policies and procedures. Our approach aims to inform scientists about data handling practices, enabling them to make informed decisions and protect sensitive information. Ongoing user studies with scientists are underway to evaluate the framework's usability, trustworthiness, and effectiveness in tackling real-world privacy challenges.

CCS Concepts

• **Security and privacy → Human and societal aspects of security and privacy**; **Privacy protections**.

Keywords

Confidential data detection, Privacy management, Privacy policies, User study, Large language models

ACM Reference Format:
Yashothara Shanmugarasa, Shidong Pan, Ming Ding, Dehai Zhao, and Thierry Rakotoarivelo. 2025. Privacy Meets Explainability: Managing Confidential Data and Transparency Policies in LLM-Empowered Science. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3706599.3720099

1 Introduction

Artificial Intelligence (AI) represents a transformative technology, revolutionizing various industries by automating complex tasks and providing intelligent solutions. A prominent development in this field is the rise of pre-trained Large Language Models (LLMs), which have set new standards in natural language processing (NLP), enabling machines to generate human-like text and perform sophisticated tasks with remarkable accuracy [17]. LLMs have become integral to scientific workflows, empowering researchers to efficiently query, analyze, and synthesize vast datasets using LLM-powered systems [10]. The 2024 Nobel Prizes in Biology and Physics, awarded for advancements in machine learning, highlight the growing role of AI and LLMs in driving scientific innovation [2]. In particular, the development of LLM agents capable of accessing external tools, making autonomous decisions, and performing multiple functions, has further enhanced the application of AI in scientific research.

However, scientists frequently work with confidential and intellectual property (IP) data belonging to their organization or individuals, which raises significant privacy concerns. These concerns stem from using LLM applications that may share sensitive information with service providers, often without a clear understanding of the associated risks, company policies on IP data protection, or the potential involvement of third-party tools. LLMs have increased data exposure compared to traditional software applications (typically require specific predefined fields for service access) as LLM prompts can include unrestricted information, and their conversational nature often encourages users to disclose more than originally intended [55]. LLM systems are fed diverse types of information from various sources in their prompts, potentially revealing more contextual data beyond the direct sensitive data in the prompts. For these reasons, many tech companies have implemented restrictions on LLM tools. For example, Samsung banned the use of tools like ChatGPT following an internal data incident [44].

Recently, the growing regulatory focus on personal data, particularly Personally Identifiable Information (PII), through frameworks such as GDPR and guidelines from organizations like the National Institute of Standards and Technology (NIST), along with decisions by government agencies like the US National Science Foundation and the Italian government [38, 39], has spurred research efforts in privacy and confidentiality preservation [21, 34, 37]. These studies

[*] Corresponding Author
[†] Shidong.Pan@anu.edu.au. Shidong Pan is also with School of Computing, Australian National University.

CHI EA '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1395-8/2025/04
https://doi.org/10.1145/3706599.3720099

aim to protect personal data within prompts by identifying and highlighting PII or encrypting them in various formats. However, while these efforts primarily focus on detecting and protecting PII, leaking confidential data extends far beyond PII when using LLMs in scientific contexts. This includes proprietary and IP data critical to organizations, such as gene or protein sequences, material names, chemical formulations, and algorithms. These types of data are highly sensitive for organizations but do not fall under the PII category, a gap that remains largely unaddressed in existing research. Moreover, LLM agents enable seamless integration with multiple external tools, allowing users to perform various tasks effortlessly through a single prompt. However, this convenience also increases the vulnerability of LLM platforms, as many users (e.g., scientists) remain unaware that their data traverses multiple pathways beyond the LLM service providers alone. This aspect of data vulnerability has not been adequately addressed in existing literature, particularly the need to inform users about the involvement of external tools, their privacy policies, and their compliance with the organization's internal policies.

To address the privacy and ethical concerns of using LLM-empowered AI tools, we propose "DataShield – Explainable Shield for Confidential Data," framework. Our approach includes three key components: i) a confidential data detection module, ii) a policy summarization module that creates clear "privacy nutrient labels" based on the privacy policies of external tools in LLM-empowered systems while also summarizing the organization's internal policies, and iii) a visualization dashboard to display data flow, triggered actions, and recommendations. Our research goal is to inform users better, particularly research scientists handling confidential information about the potential risks to their data when interacting with LLM-based tools. Our approach identifies what types of confidential data may be inadvertently exposed and whether such exposure violates internal company policies, such as the code of conduct, especially when third-party external tools are involved. The code of conduct outlines the dos and don'ts regarding data handling, confidentiality, and privacy. By detecting potential data leaks and analyzing relevant privacy policies, our framework ensures that users are fully aware of how their data is handled and empowered to make informed decisions.

We plan to conduct user studies to evaluate scientists' perceptions of the desirability, trustworthiness, and suitability of our proposed framework in addressing their privacy and ethical concerns when using LLM systems. Notably, we will use synthetic data as the hypothetical "confidential" data and open-sourced localized LLMs (e.g., Claude3-Opus) to guarantee the integrity of our user study.

1.1 Research Questions

The primary research question guiding this study is: *How can LLM-powered AI tools for scientific discovery (particularly in genomics) be designed to enhance confidentiality, ethical compliance, and ensuring alignment with organizational standards and minimizing confidential data leakage?* To address this, we propose a framework with multiple components, each investigating the following sub-research questions:

(1) **Confidential Data Exposure in LLM Interactions:** Scientists interacting with LLM-powered AI tools may inadvertently disclose confidential data, such as gene and protein names, without realizing potential risks. Existing AI systems lack mechanisms to detect and mitigate such leaks while maintaining a seamless workflow.
- **RQ1:** How can we systematically identify and classify confidential data (e.g., gene and protein names) in scientist-LLM interactions while assessing its sensitivity to provide adaptive risk-based alerts?
- **RQ2:** How can integrating user feedback and domain-specific contextual knowledge enhance the accuracy and adaptability of confidential data detection?

Our Solution: Our framework incorporates a 'confidential data detection' module that automatically identifies and categorizes sensitive terms in scientist-LLM interactions.

(2) **Lack of Policy Awareness and Compliance Support:** Scientists often struggle with navigating internal and external privacy policies when using multiple LLM-powered AI tools.
- **RQ3:** How can we effectively summarize and present internal compliance policies and external tools' privacy policies (e.g., genome sequencing tools) to scientists for improved awareness and decision-making?
- **RQ4:** What interactive mechanisms can align external privacy policies with organizational internal policies to support users in proactively managing data-sharing risks?

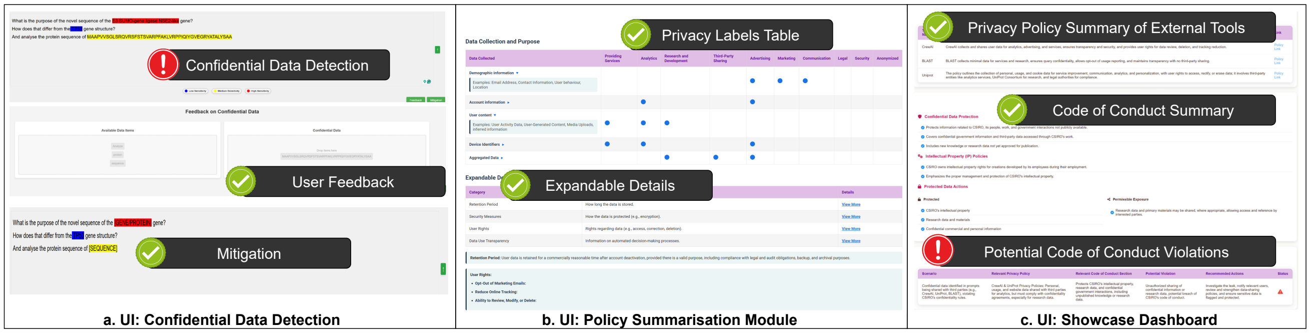
Our Solution: Our 'policy summarization' module extracts and summarizes key policy elements.

(3) **Limited Transparency in AI-Driven Data Processing:** The lack of visibility into how LLM systems process user data, trigger actions, and involve third-party tools reduces trust and usability.
- **RQ5:** What visualization techniques can enhance user comprehension of data flow, triggered actions, and third-party tool interactions within LLM-powered systems?
- **RQ6:** How does interactive visualization impact scientists' trust, usability perception, and decision-making regarding data confidentiality?

Our Solution: We introduce a 'visualization dashboard' that provides a comprehensive view of data flow, AI-triggered actions, and third-party tool interactions.

(4) **Usability-Centered Evaluation of Privacy and Explainable AI Frameworks:** Current privacy-preserving AI frameworks lack systematic evaluation methods that assess usability, compliance effectiveness, and trust.
- **RQ7:** How do scientists perceive the usability, trustworthiness, and desirability of the proposed "DataShield" framework in mitigating privacy risks?
- **RQ8:** What qualitative and quantitative evaluation methods can assess the effectiveness of the "DataShield" framework in balancing usability, compliance, and confidentiality concerns?

Our Solution: We conduct a mixed-method evaluation, combining a quantitative assessment of each component with user studies involving scientists.


What is the purpose of the novel sequence of the E3 SUMO-gene ligase NSE2-like gene?
How does that differ from the TP53 gene structure?
And analyse the protein sequence of MAAPVVSGLSRQVRSFSTSVARPFAKLVRPPIQIYGVEGRYATALYSAA
Confidential Data Detection
Low Sensitivity
Medium Sensitivity
High Sensitivity
Feedback
Mitigation
Feedback on Confidential Data
Available Data Items
Confidential Data
User Feedback
Mitigation
Privacy Labels Table
Data Collection and Purpose
Expandable Details
Privacy Policy Summary of External Tools
Code of Conduct Summary
Potential Code of Conduct Violations

a. UI: Confidential Data Detection b. UI: Policy Summarisation Module c. UI: Showcase Dashboard

Figure 1: The overview of “DataSheild" framework with three modules output: User Interface: Clear version can be found here

Figure 1 serves as a teaser image, presenting the user interface (UI) representation of our “DataShield" framework and providing an engaging overview of our contributions, which are elaborated in the subsequent sections of the paper.

2 Related Works

2.1 Studies on Personally Identifiable Information Detection

Several research efforts and commercial products are available as plugins to identify PII data in user prompts when interacting with LLM-empowered applications. PII detection methods typically use Named Entity Recognition (NER), predefined policies, LLMs, regular expressions, rule-based logic, or checksums to identify and remove sensitive details across multiple languages and contexts [13, 21, 32, 34, 35, 37]. After detecting PII data, mitigation strategies often replace sensitive information with placeholders. For example, EmojiCrypt [34] encrypts data using emojis and math operations, while others substitute or mask data [13]. Secure prompt templates and redaction tools also help minimize data leakage [35]. Hartmann et al. [23] suggested obfuscating sensitive queries with high-level descriptions, new problems, or placeholders. TextObfuscator [56] similarly obfuscates sensitive words while retaining functional meaning.

However, existing approaches focus on PII data, leveraging advanced NER taggers (e.g., spaCy NER [1]) and LLMs trained on PII datasets, making personal data detection relatively straightforward. However, detecting confidential data, such as gene names in genomics or chemical names in manufacturing, is more complex and underexplored. Additionally, high-risk scenarios where LLMs infer sensitive information from seemingly insignificant data are often overlooked [30]. Our framework addresses these gaps by focusing on confidential data detection and indirect data exposure.

2.2 Studies on Policy Summarization and Privacy Nutrition Labels

Privacy policies are legally binding documents for organizations to disclose data collection practices, mandated by privacy laws [16]. However, their complexity and length make them difficult for users to comprehend [36]. This challenge has driven extensive research efforts in policy analysis and summarization, particularly for web and mobile applications. Early studies, such as OPP-115 [54], focused on taxonomy creation and manual annotation of privacy policies, while later datasets like Bui et al. [11] introduced large-scale corpora for automated analysis. With advancements in NLP, AI-based tools such as Privee [57], Polisis [22], PolicyLint [4], PurPliance [12], and OVRseen [52] have emerged using classifiers to extract data types and entities in the privacy policies. PolicyGPT [51], based on LLMs, categorizes clauses into predefined classes. However, these tools often lack connections between extracted entities, such as which data type is collected, by whom, and for what purpose. PoliGraph [16] addressed this by using a knowledge graph to capture relationships between data types, entities, and purposes but produced outputs too complex for users. Our two-layer approach builds on PoliGraph, combining its accuracy with LLM-based summarization to produce concise, user-friendly outputs. This method avoids hallucinations and enhances readability compared to standalone LLMs.

Kelley et al. [28] introduced privacy labels inspired by nutrition labels to simplify privacy policy presentation, offering users a clear overview of data collection, usage, and sharing practices. Pan et al. [40] proposed a framework for generating privacy nutrition labels from applications' privacy policies. We also generate privacy nutrition labels from policies but extend the approach by mapping relationships between entities and purposes using PoliGraph to improve user understanding of data collection reasons. Unlike previous approaches that focus on a single policy, our method summarizes privacy policies across multiple tools to create an overall privacy label. We also analyze the company's internal policies and their alignment with external tools' data collection behaviour in their privacy policies, ensuring users can safely use LLMs while maintaining compliance.

2.3 Studies on Privacy Dashboards

Privacy dashboards can provide access to personal data in a structured and interactive manner [8]. Server-side tools for privacy settings [19, 41] dominate the market by providing a dashboard that allows users to handle privacy settings. Various approaches aim to raise privacy awareness by visualizing data collection, including privacy dashboards in [29] and [26], which use transaction logs and maps to highlight data by category and purpose. Provenance-based tools [3, 5, 8, 42] visualize data flows and sharing details. Other toolkits, such as [5, 20, 43, 47], align with GDPR transparency principles, offering more automated and adaptive use of transparency information. However, we found no existing approach that combines all involved external tools in an LLM-powered platform for privacy information visualization.

2.4 Studies on Explainability and Human Involvement in Responsible LLMs

To address the gap in model-centered research that lacks user perspectives, recent studies have explored explainability and human involvement to better understand key privacy risks and how existing research meets user needs. Zhang et al. [55] conducted a human-centered study on user disclosure behaviors and risk perceptions in LLM-based conversational agents. By analyzing sensitive disclosures in ChatGPT conversations and interviewing 19 users, they found that users face trade-offs between privacy, utility, and convenience. However, users' erroneous mental models and dark patterns in system design limited privacy awareness, while human-like interactions encouraged more sensitive disclosures, complicating these trade-offs. Similarly, studies [6, 7, 27, 33] examine the responsible integration of LLMs into research workflows. They outline a research agenda on using LLMs as research tools, addressing open empirical and ethical evaluation questions. These works explore transparency and responsible AI, focusing on issues such as perceived lack of control, distributed responsibility in the LLM supply chain, conditional ethical engagement, and competing priorities.

There are some studies [45, 48] explore black-box models and their decision-making processes, which fall outside our scope; so we do not delve further into this area. As noted in [18, 25], algorithmic transparency alone is insufficient for AI explainability, which extends beyond merely "opening" the black box. We address explainability by ensuring transparency in confidential data involvement, tool interactions, and policy alignment. While most studies identify requirements or present position papers through user studies, we found no other work combining technical contributions with user studies to evaluate them.

2.5 Our Unique Contributions

This paper makes a unique contribution by focusing on the privacy, confidentiality, and explainability aspects of LLM-empowered AI systems from the perspective of research scientists and organizations. We are the first study to address three key components to enhance the safety, ethical management, and explainability of these systems, to prevent confidential data leakage and to improve user trust and decision-making. Our approach integrates LLM-based automation, a human-in-the-loop model, and extends privacy considerations to confidentiality. We also conduct user studies with research scientists on confidentiality and compare our work to existing approaches.

Within these components, our contributions are distinctive. For confidential data detection, we focus on sensitive data like gene and protein names, which presents challenges that prevent us from relying solely on existing NER taggers or LLMs. In the second component, we use a two-layer summarization approach to maintain entity connections while ensuring readability, and our method differs from others by aligning external tool usage with the organization's internal policies through the summarization of its code of conduct. The third component features a dashboard to provide users with an overview of privacy, displaying detected confidential data, privacy policies, and compliance with the organization's internal codes of conduct. Furthermore, we conduct user studies with scientists working with such data, gathering their feedback to evaluate the desirability, trustworthiness, and suitability of our approach in addressing privacy and ethical concerns in LLM systems.

3 Proposed solution

We propose "DataShield – Explainable Shield for Confidential Data," a comprehensive framework that includes a confidential data detection module (to identify confidential information), a policy summarization module (to condense complex privacy policies of external tools into a clear "privacy nutrient label" and summarize the code of conduct of companies), and a visual dashboard (displaying data flow, triggered actions, and recommendations), especially aimed at scientists. The workflow diagram of the overall framework is shown in Figure 2, consisting of three main modules and a user study component to evaluate our system.

3.1 A Confidential Data Detection Module

The "DataShield" process begins with the confidential data detection module (RQ1, RQ2), which analyzes user prompts to identify potentially confidential data (currently focused on gene and protein data). For example, consider the following prompt provided by a biologist: "What is the purpose of the novel sequence of the *E3 SUMO-gene ligase NSE2-like gene*? And analyze this protein sequence '*MAAPVVSGLSRQVRSFSTSVARPFAKLVRPPIQIYGVEGRYATALYSAA*'." This prompt contains confidential information, such as the gene name (E3 SUMO-gene ligase NSE2-like gene) and the protein sequence. The module is designed to detect and flag such confidential data when entered into the system (Figure 1: a).

Recognizing the complexities of natural language and the potential for variations in phrasing and terminology, the module incorporated three complementary techniques. Firstly, rule-based pattern recognition uses predefined rules to detect specific phrases and data formats, identifying confidential protein sequence data based on its pattern. To enhance the module's ability to handle more nuanced language and contextual information, a Retrieval Augmented Generation (RAG) approach was integrated. This approach involved utilizing LLM to extract relevant information from a knowledge base containing a collection of public gene and protein names sourced from the Uniprot library [14]. This enabled the system to generate contextually relevant responses and enhance the accuracy of identifying confidential data by incorporating gene and protein names from the knowledge base, providing more accurate information than relying solely on the LLM, as it learns from these examples. Finally, we incorporated a human-in-the-loop approach to define additional confidential data as specified by the company or individual. The system then checks these user-defined data against prompts using fuzzy logic searching, enabling it to identify potential matches even with minor discrepancies.

In addition to detecting confidential data, we incorporate contextual information, such as well-cited genes or proteins, novelty

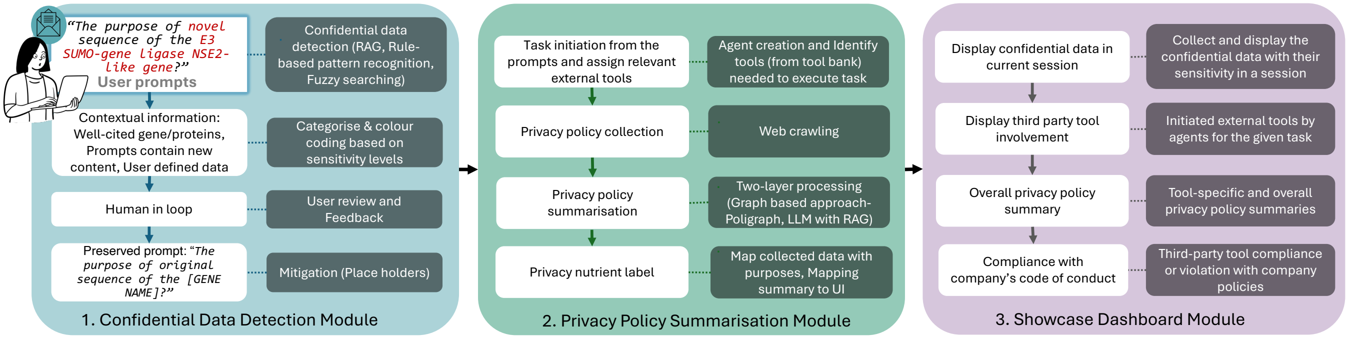

"The purpose of novel sequence of the E3 SUMO-gene ligase NSE2-like gene?"
User prompts
Confidential data detection (RAG, Rule-based pattern recognition, Fuzzy searching)
Contextual information: Well-cited gene/proteins, Prompts contain new content, User defined data
Categorise & colour coding based on sensitivity levels
Human in loop
User review and Feedback
Preserved prompt: "The purpose of original sequence of the [GENE NAME]?"
Mitigation (Place holders)
1. Confidential Data Detection Module
Task initiation from the prompts and assign relevant external tools
Agent creation and Identify tools (from tool bank) needed to execute task
Privacy policy collection
Web crawling
Privacy policy summarisation
Two-layer processing (Graph based approach- Poligraph, LLM with RAG)
Privacy nutrient label
Map collected data with purposes, Mapping summary to UI
2. Privacy Policy Summarisation Module
Display confidential data in current session
Collect and display the confidential data with their sensitivity in a session
Display third party tool involvement
Initiated external tools by agents for the given task
Overall privacy policy summary
Tool-specific and overall privacy policy summaries
Compliance with company's code of conduct
Third-party tool compliance or violation with company policies
3. Showcase Dashboard Module

Figure 2: Workflow Diagram of "DataShield" – Explainable Shield for Confidential Data (RAG: Retrieval-Augmented Generation)

exposure in the prompt, and user-defined confidential information, to categorize the severity of the data. The colour-code scheme is as follows: red for the high sensitivity, yellow for the medium sensitivity, and blue for the low sensitivity. This design choice is based on its intuitive and widely recognized use in risk assessment [24], allowing scientists to gauge the level of confidentiality associated with their input immediately.

Additionally, we have addressed indirect data exposure through prompts, where the inference capabilities of LLMs can reveal sensitive information that is not explicitly mentioned in the prompt. For instance, in the prompt, "We have identified Gene_B in a wild maize relative, sharing a conserved domain with an Arabidopsis receptor involved in salt stress signalling. How can we annotate and validate its role in salinity tolerance?", although Gene_B is unnamed, references to salt stress and conserved domains imply potential candidates like SOS1 or related SOS pathway components, highlighting the risk of indirect data exposure.

The next feature of this module allows users to provide feedback on the detected confidential data, enabling them to mark it as confidential or not, with the system learning from this feedback. Furthermore, we propose a simple mitigation technique using placeholders (e.g., [GENE_NAME]) to redact confidential data from the prompt.

3.2 A Policy Summarization Module

The second core component of "DataShield" is a Policy Summarization Module (RQ3, RQ4), which aims to provide users with clear and concise summaries of complex privacy policies from external tools and internal company policies, such as the code of conduct (i.e., outlining what scientists can do to comply with company guidelines), thereby making it easier for them to understand the implications of sharing their data. The process begins by identifying relevant external tools from the tool bank, including 40 CrewAI tools [15], LangChain [31], and gene-related tools [9]. These tools were particularly aligned with the genomics use case scenario we focused on. To identify the potential tools, we created an agentic environment based on the user prompt, which helped us identify the most appropriate tools for executing tasks specified by the users. After determining the relevant tools, we then scraped the privacy policies associated with each of these tools online. The privacy policies collected for each tool are then processed using a two-layer summarization approach. The first layer employed a graph-based method, PoliGraph [16], to analyze the structure and extract privacy practice disclosures from privacy policies. This approach mapped relationships between identified data entities, data types, collection purposes, and policy procedures into a structured format. However, the output of PoliGraph, while comprehensive, is not in a human-readable format and is often lengthy. A second layer was introduced to further process the extracted information by utilizing an LLM with the RAG strategy. This layer complemented the graph-based analysis by extracting important policy information, identifying important clauses and conditions, and generating concise, meaningful summaries that highlighted essential aspects of the policies. To enhance the readability further, we transformed the extracted content in the form of "privacy nutrient labels". The labels provide a concise summary of the most critical elements of the privacy policy, including data types, purposes, retention periods, security measures, user rights, and third-party involvement. By presenting this information in an intuitive format (as shown in Figure 1: b), users could quickly understand the key implications of each policy, promoting transparency.

Similarly, we leveraged another LLM empowered by RAG to summarize the company's internal policies, resulting in insights such as Confidential Data, IP Policies, Protected vs. Exposed Information, Violations of Confidentiality and IP data, and additional policies required for compliance. Using the LLM, we assessed the summarized privacy policies of external tools and the company's code of conduct to scrutinize compliance and potential violations (Figure 1: c).

3.3 A Visualization Dashboard Module

The Visualization Dashboard Module (RQ5, RQ6) (Figure 1: c), served as the central interface for presenting the results of the confidential data detection and summarization processes with a clear and intuitive view. The dashboard displays the confidential data detected highlighting its sensitivity level. It also provides information about any relevant external tools involved in the prompt execution, along with individual tool privacy nutrient labels. Critically, the dashboard displayed the overall privacy policy summaries generated by the Policy Summarization Module, along with their compliance/violation with summarized internal company policies, providing users with a comprehensive understanding of the applicable privacy practices and compliance requirements. This integration of information within a single dashboard aimed to empower users with the knowledge they needed to make informed decisions about their data.

3.4 User Study

A user study (RQ7, RQ8) will be conducted to evaluate the effectiveness and usability of the integrated framework by interacting with it. We obtained ethical approval from our Human Research Ethics Committee. Participants in this study would be composed by scientists from various subjects, such as genomics, computer science, and chemical and material engineering, who utilize LLM-empowered AI tools in organizations. To expand beyond genomics, we will enhance the confidential data detection module to accommodate confidential data from other domains. We anticipate 30 to 40 scientists will participate in the user study. Participants will receive a link to the “DataShield" framework, a questionnaire, and a consent form with privacy information. The questionnaire is designed to ensure that no personally identifiable information is collected. The study consists of three sections: 1) demographic questions assessing familiarity with LLM-empowered tools and privacy concerns; 2) hands-on experience with “DataShield" using a synthetic dataset (will be given to the users) containing prompts that a scientist might ask; and 3) a user experience survey on how “DataShield" addresses user concerns. The synthetic dataset, created with ChatGPT-4o using publicly known gene or material names and synthetic company names, contains no real or confidential data, ensuring no privacy risks. We use a prompt in ChatGPT to create the synthetic dataset. This approach ensures the integrity of the synthetic dataset while focusing on testing privacy management mechanisms effectively. Humans will review every prompt ChatGPT generates to ensure it is produced as expected and does not unintentionally leak any confidential information. To enhance participant security, they will interact with DataShield, powered by Claude3-Opus localized, open-source LLM—instead of an online LLM application. Claude3-Opus serves as the backend for the DataShield framework, offering a secure and flexible environment while ensuring that data privacy is maintained throughout the testing process. Participants are asked to use the “DataShield” to simulate their daily usage scenario for 30 minutes.

Following the hands-on experience, participants will be surveyed on several key aspects, including the desirability of the system, the level of trust users placed in its accuracy, the perceived information load, the accuracy of the information provided, overall user satisfaction, and suggestions for improvement. The data gathered from this user study provided valuable insights into the system's strengths and weaknesses, informing further refinements and improvements.

4 Preliminary Results

This research utilizes a mixed-methods approach, combining qualitative methods (through user studies) and quantitative techniques to evaluate the effectiveness of our framework, both overall and at the level of individual modules.

The effectiveness of the confidential data detection module was quantitatively assessed using accuracy by comparing it with baseline biomedical NER tools such as BERN2 [50], HunFlair1 [53], and HunFlair2 [46], as well as various LLMs, including GPT-4o, and local LLMs such as Mistral-Large-2407, Claude3-Opus, Claude3-Sonnet, Llama-3.1-70b, and Claude2. We used 500 test sentences from the BC2GM dataset [49], which was used to train the BERN2 model. The results, summarized in Table 1, show that while GPT-4o excelled in accuracy, its precision and F1 scores were lower. The BERN2 model yielded strong performance; However, these results may not be generalized well to other scenarios, as the test dataset was the same one used to train the BERN2 model. Future work will evaluate performance across additional public datasets to ensure generalizability across diverse models and scenarios. For the user study experiments, we selected Claude3-Opus due to safety and privacy considerations. Although participants are advised to interact with synthetic datasets during testing, Claude3-Opus was chosen as an additional precaution to mitigate risks in the event of inadvertent disclosure of confidential data. Claude3-Opus achieved reasonable accuracy, precision, recall, and F1 scores, balancing safety and effectiveness. Moreover, organizations can avoid relying on external LLM service providers to detect confidential data by deploying a small local LLM on their premises. This approach enhances confidentiality in handling highly sensitive data and reduces maintenance costs by utilizing a compact local model to process data before sharing them with external providers.

Table 1: Performance comparison of various tools for confidential data detection. (RAG: Retrieval Augmented Generation)

Tool	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
BERN2	90.01	79.06	90.01	84.18
HunFlair1	74.19	93.80	74.19	82.85
HunFlair2	70.38	91.31	70.38	79.49
GPT-4o + RAG	97.56	69.78	97.56	81.36
Mistral-Large-2407 + RAG	75.08	90.19	75.15	81.98
Claude3-Opus + RAG	76.73	94.10	76.80	84.57
Claude3-Sonnet + RAG	65.59	92.73	65.59	76.83
Llama-3.1-70b + RAG	68.29	89.85	68.29	77.60
Claude2 + RAG	47.99	84.75	47.99	61.28

The outputs of the Policy Summarization Module and the Visualization Dashboard are illustrated in Figure 1: c,d. The Policy Summarization Module will be quantitatively evaluated using a question-answer approach by deriving a set of questions related to the policy, such as the purpose of data collection and retention periods. Two reviewers will assess the entire policy and its summary to determine how often the summarization provides correct answers based on their agreement. The Visualization Dashboard will be evaluated through a user study, which is currently underway.

5 Discussion and Future Work

This paper addresses the confidentiality and ethical management challenges of LLM-powered AI tools for scientific discovery from the perspective of scientists. In future work, We will evaluate our approach through user studies with scientists, combining qualitative insights and quantitative assessments. The confidential data detection module will be assessed for accuracy across data domains, and policy summarization will be tested using a question-answer approach and LLM-based evaluations. We also aim to extend the framework to material and data science for broader applicability. Future work will also enhance system reliability by mitigating hallucinations using strategies like prompt engineering and reinforcement learning from human feedback. This aligns with our human-in-the-loop framework, ensuring real-time feedback and protection of sensitive data, thereby promoting the safe adoption of LLMs in scientific research. We will also explore DataShield's scalability in multi-user environments to evaluate its performance in collaborative research contexts.

References

[1] [n. d.]. EntityRecognizer · spaCy API Documentation — spacy.io. https://spacy.io/api/entityrecognizer. [Accessed 23-01-2025].

[2] Nobel Prize Outreach 2025. 2024. The Nobel Prize in Physics and Biology 2024. https://www.nobelprize.org/prizes/physics/2024/summary/.

[3] Rocio Aldeco Perez and Luc Moreau. 2008. Provenance-based auditing of private data use. *Visions of Computer Science - BCS International Academic Conference (VOCS)* (2008).

[4] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association, Santa Clara, CA, 585–602. https://www.usenix.org/conference/usenixsecurity19/presentation/andow

[5] Julio Angulo, Simone Fischer-Hübner, Tobias Pulls, and Erik Wästlund. 2015. Usable Transparency with the Data Track: A Tool for Visualizing Data Disclosures. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI EA '15)*. Association for Computing Machinery, New York, NY, USA, 1803–1808. doi:10.1145/2702613.2732701

[6] Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Tangi Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. 2024. LLMs as Research Tools: Applications and Evaluations in HCI Data Work. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 479, 7 pages. doi:10.1145/3613905.3636301

[7] Kristian Gonzalez Barman, Nathan Wood, and Pawel Pawlowski. 2024. Beyond transparency and explainability: on the need for adequate and contextualized user guidelines for LLM use. *Ethics and Information Technology* 26, 3 (2024), 47.

[8] Christoph Bier, Kay Kühne, and Jürgen Beyerer. 2016. PrivacyInsight: The Next Generation Privacy Dashboard. In *Privacy Technologies and Policy*, Stefan Schiffner, Jetzabel Serna, Demosthenes Ikonomou, and Kai Rannenberg (Eds.). Springer International Publishing, Cham, 135–152.

[9] Jean-Simon Brouard, Flavio Schenkel, Andrew Marete, and Nathalie Bissonnette. 2019. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. *Journal of animal science and biotechnology* 10 (2019), 1–6.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[11] Duc Bui, Kang G Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies* (2021).

[12] Duc Bui, Yuan Yao, Kang G Shin, Jong-Min Choi, and Junbum Shin. 2021. Consistency analysis of data-usage purposes in mobile apps. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2824–2843.

[13] Yu Chen, Tingxin Li, Huiming Liu, and Yang Yu. 2023. Hide and Seek (HaS): A Lightweight Framework for Prompt Privacy Protection. arXiv:2309.03057 [cs.CR]

[14] UniProt Consortium. 2015. UniProt: a hub for protein information. *Nucleic acids research* 43, D1 (2015), D204–D212.

[15] CrewAI. [n. d.]. Tools - CrewAI — docs.crewai.com. https://docs.crewai.com/concepts/tools. [Accessed 23-01-2025].

[16] Hao Cui, Rahmadi Trimananda, Athina Markopoulou, and Scott Jordan. 2023. {PoliGraph}: Automated privacy policy analysis using knowledge graphs. In *32nd USENIX Security Symposium (USENIX Security 23)*. 1037–1054.

[17] Travis Dyde. 2023. Documentation on the emergence, current iterations, and possible future of Artificial Intelligence with a focus on Large Language Models. (2023).

[18] Upol Ehsan, Elizabeth A Watkins, Philipp Wintersberger, Carina Manger, Sunnie SY Kim, Niels Van Berkel, Andreas Riener, and Mark O Riedl. 2024. Human-centered explainable AI (HCXAI): Reloading explainability in the era of large language models (LLMs). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[19] Google. [n. d.]. Sign in - Google Accounts — google.com. https://www.google.com/dashboard/. [Accessed 23-01-2025].

[20] Elias Grünewald and Frank Pallas. 2021. TILT: A GDPR-Aligned Transparency Information Language and Toolkit for Practical Privacy Engineering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 636–646. doi:10.1145/3442188.3445925

[21] David Haber. 2024. Introducing Lakera Guard – Bringing Enterprise-Grade Security to LLMs with One Line of Code | Lakera – Protecting AI teams that disrupt the world. — lakera.ai. https://www.lakera.ai/blog/lakera-guard-overview. [Accessed 15-05-2024].

[22] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*. 531–548.

[23] Florian Hartmann, Duc-Hieu Tran, Peter Kairouz, Victor Cărbune, and Blaise Aguera y Arcas. 2024. Can LLMs get help from other LLMs without revealing private information? arXiv:2404.01041 [cs.LG] https://arxiv.org/abs/2404.01041

[24] Roger C Jensen, Royce L Bird, and Blake W Nichols. 2022. Risk assessment matrices for workplace hazards: Design for usability. *International journal of environmental research and public health* 19, 5 (2022), 2763.

[25] Naman Kandhari, Bharat Tripathi, Sheetesh Kumar, Kulvinder Singh, and Narendra Pal Singh. 2024. Responsible AI Framework For Large Language Models (LLMs). In *2024 11th International Conference on Advances in Computing and Communications (ICACC)*. IEEE, 1–6.

[26] Elahe Kani-Zabihi and Martin Helmhout. 2012. Increasing service users' privacy awareness by introducing on-line interactive privacy features. In *Information Security Technology for Applications: 16th Nordic Conference on Secure IT Systems, NordSec 2011, Tallinn, Estonia, October 26-28, 2011, Revised Selected Papers 16*. Springer, 131–148.

[27] Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2024. " I'm categorizing LLM as a productivity tool": Examining ethics of LLM use in HCI research practices. *arXiv preprint arXiv:2403.19876* (2024).

[28] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. 2009. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (Mountain View, California, USA) *(SOUPS '09)*. Association for Computing Machinery, New York, NY, USA, Article 4, 12 pages. doi:10.1145/1572532.1572538

[29] Jan Kolter, Michael Netter, and Günther Pernul. 2010. Visualizing past personal data disclosures. In *2010 International Conference on Availability, Reliability and Security*. IEEE, 131–139.

[30] Jacob Leon Kröger, Leon Gellrich, Sebastian Pape, Saba Rebecca Brause, and Stefan Ullrich. 2022. Personal information inference from voice recordings: User awareness and privacy concerns. *Proc. Priv. Enhancing Technol.* 2022, 1 (2022), 6–27.

[31] LangChain. 2024. Tools |LangChain — python.langchain.com. https://python.langchain.com/v0.1/docs/modules/tools/. [Accessed 23-01-2025].

[32] Tianshi Li, Sauvik Das, Hao-Ping Lee, Dakuo Wang, Bingsheng Yao, and Zhiping Zhang. 2024. Human-Centered Privacy Research in the Age of Large Language Models. *arXiv preprint arXiv:2402.01994* (2024).

[33] Q Vera Liao and Jennifer Wortman Vaughan. 2023. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941* 10 (2023).

[34] Guo Lin, Wenyue Hua, and Yongfeng Zhang. 2024. EmojiCrypt: Prompt Encryption for Secure Communication with Large Language Models. arXiv:2402.05868 [cs.CL] https://arxiv.org/abs/2402.05868

[35] Aatish Mandelecha. 2024. How to Secure Sensitive Data in LLM Prompts? — strac.io. https://www.strac.io/blog/secure-sensitive-data-in-llm-prompts. [Accessed 14-05-2024].

[36] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *Isjlp* 4 (2008), 543.

[37] Microsoft. 2024. Home - Microsoft Presidio — microsoft.github.io. https://microsoft.github.io/presidio/. [Accessed 21-01-2025].

[38] National Science Foundation (NSF). [n. d.]. Notice to research community: Use of generative artificial intelligence technology in the NSF merit review process. https://new.nsf.gov/news/notice-to-the-research-community-on-ai. [Accessed 21-01-2025].

[39] Institute of Standards and Technology (NIST). 2024. National Institute of Standards and Technology — nist.gov. https://www.nist.gov/. [Accessed 21-01-2025].

[40] Shidong Pan, Thong Hoang, Dawen Zhang, Zhenchang Xing, Xiwei Xu, Qinghua Lu, and Mark Staples. 2023. Toward the cure of privacy policy reading phobia: Automated generation of privacy nutrition labels from privacy policies. *arXiv*

preprint arXiv:2306.10923 (2023).

[41] PRIME. 2004. PRIME - Privacy and Identity Management for Europe — Portal for the PRIME Project — prime-project.eu. https://prime-project.eu/. [Accessed 23-01-2025].

[42] Tobias Pulls, Roel Peeters, and Karel Wouters. 2013. Distributed privacy-preserving transparency logging. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*. 83–94.

[43] Philip Raschke, Axel Küpper, Olha Drozd, and Sabrina Kirrane. 2018. Designing a GDPR-compliant and usable privacy dashboard. *Privacy and Identity Management. The Smart Revolution: 12th IFIP WG 9.2, 9.5, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Ispra, Italy, September 4-8, 2017, Revised Selected Papers 12* (2018), 221–236.

[44] Siladitya Ray. 2023. Samsung Bans ChatGPT Among Employees After Sensitive Code Leak — forbes.com. https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/. [Accessed 24-01-2025].

[45] Walid S Saba. 2023. Towards explainable and language-agnostic LLMs: symbolic reverse engineering of language at scale. *arXiv preprint arXiv:2306.00017* (2023).

[46] Mario Sänger, Samuele Garda, Xing David Wang, Leon Weber-Genzel, Pia Droop, Benedikt Fuchs, Alan Akbik, and Ulf Leser. 2024. HunFlair2 in a cross-corpus evaluation of biomedical named entity recognition and normalization tools. *Bioinformatics* 40, 10 (2024), btae564.

[47] Marija Schufrin, Steven Lamarr Reynolds, Arjan Kuijper, and Jorn Kohlhammer. 2020. A visualization interface to improve the transparency of collected personal data on the internet. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1840–1849.

[48] Ishika Shruti, Amol Kumar, Arjun Seth, et al. 2024. Responsible Generative AI: A Comprehensive Study to Explain LLMs. In *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET*. IEEE, 1–6.

[49] Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of BioCreative II gene mention recognition. *Genome biology* 9 (2008), 1–19.

[50] Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* 38, 20 (2022), 4837–4839.

[51] Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xiang Li, Tianming Liu, and Lei Fan. 2023. PolicyGPT: Automated Analysis of Privacy Policies with Large Language Models. arXiv:2309.10238 [cs.CL] https://arxiv.org/abs/2309.10238

[52] Rahmadi Trimananda, Hieu Le, Hao Cui, Janice Tran Ho, Anastasia Shuba, and Athina Markopoulou. 2022. {OVRseen}: Auditing network traffic and privacy policies in oculus {VR}. In *31st USENIX security symposium (USENIX security 22)*. 3789–3806.

[53] Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics* 37, 17 (2021), 2792–2794.

[54] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1330–1340.

[55] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 156, 26 pages. doi:10.1145/3613904.3642385

[56] Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Yuran Wang, Yong Ding, Yibo Zhang, Qi Zhang, and Xuan-Jing Huang. 2023. Textobfuscator: Making pre-trained language model a privacy protector via obfuscating word representations. In *Findings of the Association for Computational Linguistics: ACL 2023*. 5459–5473.

[57] Sebastian Zimmeck and Steven M Bellovin. 2014. Privee: An architecture for automatically analyzing web privacy policies. In *23rd USENIX Security Symposium (USENIX Security 14)*. 1–16.