A Theoretical Framework for Ethics and Safety Assurance of Generative Artificial Intelligence in Smart Libraries

Authors: Liu Yaqi, Zheng Wenjie, Su Yueli, Liu Yaqi

Date: 2024-09-04T00:00:00+00:00

Abstract

[Purpose/Significance] With the widespread application of Generative AI (GenAI) in smart libraries, ethical and security issues have become increasingly prominent, and how to construct an effective ethical and security safeguard mechanism has become a key issue in the library field. This paper aims to explore and establish a systematic theoretical framework for ethical and security safeguards to address the ethical risks and security challenges faced by GenAI technology in smart library applications, and to promote the healthy and sustainable development of smart libraries. [Process/Method] Through literature analysis and other methods, this study reviews relevant research literature and practical overviews, analyzing the current theoretical research and practical progress of generative AI applications in smart libraries. Furthermore, it conducts an in-depth analysis of the current status and ethical security issues of GenAI applications in smart libraries to construct an ethical and security safeguard theoretical framework. [Results/Conclusion] The ethical and security safeguard theoretical framework centers on a multi-agent collaborative governance model, which generates supplementary prompts that meet ethical requirements through multi-round feedback updates, thereby ensuring that AI-generated content (AIGC) complies with ethical demands. The proposed theoretical framework provides a systematic solution for the ethical and security management of GenAI in smart libraries, helping to ensure the ethically compliant application of GenAI technology and thus promoting the long-term development of smart libraries.

Full Text

Research on an Ethical and Security Assurance Theoretical Framework for Generative Artificial Intelligence in Smart Libraries

Liu Yaqi, Zheng Wenjie, Su Yueli
School of Information Engineering, Zhongnan University of Economics and Law, Wuhan 430073

Abstract
[**Purpose/Significance**] With the widespread application of Generative Artificial Intelligence (GenAI) in smart libraries, ethical and security issues have become increasingly prominent, making the construction of an effective ethical security assurance mechanism a critical challenge in the library field. This paper aims to explore and establish a systematic ethical security assurance theoretical framework to address the ethical risks and security challenges faced by GenAI technology in smart library applications, thereby promoting the healthy and sustainable development of smart libraries. [**Process/Methods**] Through literature analysis and other methods, this study reviews relevant research literature and practical developments, analyzing the current theoretical research and practical progress of GenAI applications in smart libraries. Furthermore, it conducts an in-depth analysis of the current status and ethical security issues of GenAI applications in smart libraries to construct an ethical security assurance theoretical framework. [**Results/Conclusions**] The ethical security assurance theoretical framework centers on a multi-agent collaborative governance model, which generates supplementary prompts that meet ethical requirements through multi-round feedback updates, thereby ensuring that Artificial Intelligence Generated Content (AIGC) complies with ethical demands. The proposed theoretical framework provides a systematic solution for the ethical security management of GenAI in smart libraries, helping to ensure the ethically compliant application of GenAI technology and thus promoting the long-term development of smart libraries.

Keywords: Smart Libraries, Generative Artificial Intelligence, AI Ethics, Multi-Agent Systems
Classification Number: G250.76

1 Introduction

The report from the 20th National Congress of the Communist Party of China explicitly proposed the strategic goal of "implementing the national cultural digitization strategy and improving the modern public cultural service system," which points the direction for China's cultural development. As an important component of the modern public cultural service system, smart libraries have been incorporated into the 14th Five-Year Plan for cultural development, becoming a key force in promoting the digital transformation of culture.

The development of smart libraries requires robust technological resources as support. Emerging technologies, particularly Generative Artificial Intelligence (GenAI), with their exceptional information generation and human-computer interaction capabilities, have empowered libraries in collection resource construction, service model innovation, and user experience upgrading, enabling the transition from digital libraries to smart libraries (Li Tao, 2024). In application scenarios such as reading promotion services, subject services, information literacy education, scientific research services, and reference consultation services, smart libraries can innovate service models through GenAI technology, providing users with immersive interactive experiences to achieve knowledge value-added.

However, while GenAI technology brings tremendous opportunities for the intelligent development of libraries, it also triggers a series of ethical risks and security challenges. In terms of moral ethics and values, GenAI may generate content that does not align with core social values, potentially containing bias and discrimination against specific groups. Such content not only affects the shaping of readers' values but may also negatively impact social harmony and stability. Regarding intellectual property rights, content generated through recombination by GenAI may neglect to cite information sources, thereby infringing on the rights of original authors and reducing the quality and credibility of library services. Furthermore, in terms of privacy protection and data security, GenAI technology may use user interaction data for training without permission, potentially leading to the leakage of readers' personal information and violating their privacy rights. Finally, regarding information quality, Artificial Intelligence Generated Content (AIGC) may suffer from inaccuracies, incompleteness, and inconsistencies, making it difficult to ensure the correctness of knowledge acquired by readers.

Faced with these challenges, ensuring that the application of GenAI technology meets ethical and security standards while promoting its deep integration with library services has become an urgent issue. The risks faced by GenAI large models used in smart libraries typically originate from the models' own security issues and potential problems during different training stages (Wang Jing et al., 2024). During the generation phase of GenAI large models, users may intentionally guide the model to generate specific content, or queries may involve personal privacy, illegal activities, bias and discrimination, and unauthorized information, all of which may cause library GenAI large models to output inappropriate content. In this process, ensuring that AI output is accurate and ethically compliant is crucial. Through refined prompt guidance of AI during the generation phase, the model's understanding of context can be enhanced, ambiguity in queries can be reduced, and the generation of inappropriate content involving privacy violations, illegal activities, bias and discrimination, and unauthorized information can be avoided. Therefore, this paper proposes an ethical security assurance theoretical framework by analyzing the current status of theoretical and applied research on GenAI in smart libraries, aiming to achieve dynamic management and technical assurance of ethical risks and promote the

sustainable development of smart libraries.

2 Applications and Challenges of Generative AI in Smart Libraries

2.1.1 Theoretical Research on AIGC Applications in Libraries

Libraries' provision of knowledge-intensive services relies on the support of information technology, and GenAI provides technical possibilities for the transformation and upgrading of library services (Zhang Hai et al., 2024). In the field of smart libraries, GenAI effectively promotes innovation in library technology and service systems (Zhang Hui et al., 2023). Scholars have focused on exploring paths to empower smart library construction with GenAI, conducting research in conjunction with key business scenarios such as library knowledge services and reading promotion (Guo Yajun et al., 2023). ChatGPT is essentially a chatbot program built on a natural language processing system that can chat and communicate like a human and perform tasks such as writing articles, press releases, and drawing according to human instructions. Some scholars have taken ChatGPT and other GenAI large models as examples to discuss the application of AIGC in empowering smart library services. On one hand, relevant research focuses on the theoretical paths (Guo Yajun et al., 2023) and opportunities and challenges (Zhou Xu, 2023) of ChatGPT empowering smart libraries. Wu Ruohang et al. (2023) summarized that ChatGPT should adhere to the core concepts of personalization, efficiency, accuracy, diversity, and equality in empowering library services, and conduct service innovation in intelligent reading services and intelligent consulting services. Li Shuning et al. (2023) tested and investigated ChatGPT from the perspective of library business, analyzing its functions and problems. Hou Zhijiang et al. (2024) summarized the main application modes of ChatGPT in library practice from three perspectives—resource intelligence, knowledge services, and business innovation—based on practical application cases, and then proposed a specific implementation path led by cognitive reconstruction, starting with prompt technology, and driven by scenario demands. On the other hand, relevant research focuses on the application of ChatGPT in single business scenarios of libraries. An Zidong et al. (2023) proposed a library literature resource management framework based on GenAI technology, covering four aspects: AIGC-driven construction, integration, reader service optimization, and resource evaluation. Gong Furong (2023) analyzed the impact of generative AI tools such as ChatGPT on digital literacy education in universities through empirical research, proposed localized education strategies, and evaluated their positive effects on students' thinking and emotions. Guo Yajun et al. (2024) explored the adaptability of large language models in empowering library reference consultation services by analyzing their generation mechanisms and application modes. Sun Shangfeng et al. explored the application of generative AI in smart reading promotion and proposed a library promotion model including four levels: resource layer, platform layer, space layer, and service layer. Unlike ChatGPT-type GenAI models, Sora and

similar GenAI tools are often used as text-to-video or image-to-video generation tools. Li Tao et al. (2024) started with the basic principles of Sora to explore its applications in smart library construction, including visual knowledge services, smart virtual space creation, and upgrades in library intelligent information construction and management. Simultaneously, they provided legal regulation suggestions for potential risks from perspectives such as clarifying work rights attribution and identifying video works.

In current theoretical research, scholars generally focus on the diverse applications of GenAI in smart libraries, emphasizing its importance in service innovation, knowledge management, and user experience enhancement, while also pointing out corresponding ethical and technical challenges.

2.1.2 Current Status of AIGC Application Practices in Libraries

Based on theoretical research on GenAI applications in smart libraries, libraries worldwide (especially university libraries) have begun active application attempts. Some libraries have applied ChatGPT and other GenAI tools to library services and related learning resources to improve service efficiency and user experience. In 2023, Zayed University Library developed a chatbot named Aisha for library services, marking the first publicly disclosed case of generative AI applied in the library field internationally (Yrjo et al., 2023). Zayed University Library built the generative chatbot Aisha to reshape library reference consultation services, providing convenient and personalized services and innovative solutions for library users while reducing the workload of librarians. In the same year, the University of Michigan built a GenAI tool service platform[1], providing free customized GenAI tool suite services to all faculty and students. Harvard University's School of Engineering and Applied Sciences collaborated with technology companies to develop a tool called "AI Sandbox"[2], which can access multiple popular generative AI chat products while ensuring data is secure and controllable and will not be collected by third parties for AI model training. In September 2023, Harvard University released guidelines for using ChatGPT and other generative AI in classrooms[3], helping teachers and students use generative AI products in safe and reliable ways. Princeton University Library built a Generative AI LibGuide[4], providing an introduction to "generative AI," ethics and copyright, AI citation methods, and extended resources in a resource navigation format, enabling faculty and students to initially understand and learn AI-related knowledge. In addition to providing the Generative AI LibGuide, Princeton University Library also conducted AI workshops to cultivate faculty and students' skills in AI research and application, helping to improve AI literacy.

In recent years, domestic universities have also begun actively exploring GenAI applications, promoting the development and practice of related technologies. The Tsinghua University Library Information Service Dialogue Robot "Qing Xiaotu" is a dialogue robot system centered on natural language processing technology, featuring a new interactive interface and interaction methods. Readers

can consult the robot via voice or text about common library issues such as book borrowing and returning, electronic resources, seat reservation, collection catalog, and in-library services, and the robot can respond via voice, text, images, and other methods. CNKI Library has developed an AI academic assistant, providing question-answering enhanced retrieval and generative knowledge services[5]. The CNKI AI-enhanced retrieval and AI academic assistant, featuring question-answering enhanced retrieval and generative knowledge services, will be applied to improve literature retrieval and reading efficiency, assist in thematic research and knowledge acquisition, and provide comprehensive support for academic research and technological innovation[6]. The China-Singapore Friendship Library launched a large model, library version of "ChatGPT" digital human Xiao Bo, introducing industry-popular large model technology. Through specialized training on a large amount of library business data, it possesses strong semantic understanding and natural language processing capabilities. When using it, readers can directly interact with digital human Xiao Bo to obtain fast, accurate, and reliable information consulting services[7]. Meanwhile, numerous domestic university libraries use the Chatlibrary service from Beijing Yingke Qianxin Technology Co., Ltd.[8] to upgrade library resource digital management, enrich user experience through intelligent retrieval and personalized recommendation systems, while providing electronic resource management and knowledge services to promote knowledge sharing and cultural dissemination. Currently, GenAI applications in various libraries are gradually expanding, not only improving service efficiency and user experience but also demonstrating their potential in reference consultation, resource management, and other aspects. In the future, GenAI will be more widely applied in libraries, and relevant application norms and ethical standards will continue to be improved to ensure their compliant and safe implementation.

GenAI applications in smart libraries have achieved significant progress at both the theoretical research and practical levels. From a theoretical research perspective, scholars have explored the potential of AIGC in library service innovation, knowledge management, and user experience enhancement, while also pointing out relevant ethical and technical challenges. At the practical level, many university libraries at home and abroad have begun applying ChatGPT and other generative AI tools to reference consultation, resource management, and other fields, significantly improving service efficiency and user experience. However, despite the increasingly widespread application of AIGC in libraries, practical research on its ethical security issues remains relatively insufficient.

2.2.1 Research on Ethical Risks of AIGC in Libraries

The application of GenAI in smart libraries has brought unprecedented opportunities and challenges for library development. As technology continues to advance, AIGC has demonstrated significant advantages in library services, such as improving service efficiency, optimizing resource allocation, and personalizing user experience. However, these technological advances are accompanied

by a series of ethical risks that have gradually revealed their potential harm in practical applications and require sufficient attention. When GenAI is applied in smart libraries, it mainly faces three types of ethical risks: output authenticity issues, personal data leakage, and intellectual property rights and academic ethics.

The authenticity of output data is one of the core ethical risks of GenAI in library applications. Companies such as Amazon, Tencent, and OpenAI have acknowledged that GenAI large models suffer from quality issues such as bias in training data due to limited high-quality datasets, rough processing, and inadequate review (Zhu Yu et al., 2023). Meanwhile, generative deceptive AI uses AI to generate realistic false data that may deceive systems or humans. Attackers can use existing data to generate false content that is difficult to identify. Technological advances have reduced the cost of producing false content and increased the likelihood of successful attacks[9]. GenAI relies on large-scale language models that generate content based on statistical probabilities of vast amounts of data, making the accuracy and authenticity of its output results often difficult to guarantee. Insufficient data, noise and randomness, training bias, and lack of contextual understanding can all trigger this issue (Liu Li et al., 2023). In practical applications, when libraries use intelligent dialogue robots to provide reference consultation and intelligent retrieval services, the generated information may be inaccurate or false. This not only affects the quality of information users obtain but may also lead to user misunderstanding. Therefore, when using these technologies, libraries must combine manual review mechanisms to examine and correct generated content to ensure information accuracy and reliability and prevent misleading users. Meanwhile, GenAI finds it difficult to determine whether user input information meets ethical requirements and may thus generate unethical responses. This issue also places higher demands on the supervision of GenAI in smart libraries.

At the same time, the risk of personal privacy data leakage is also worth noting. Smart libraries typically process large amounts of user data, including personal identity information, behavioral data, and interaction records. If these data lack strict protection measures during collection and processing, privacy leakage or data abuse may occur. In cases of poor data management, users' sensitive information may be illegally accessed, leaked, or abused, leading to security issues such as online fraud and privacy violations (Xu Fang, 2024). Therefore, when implementing AIGC technology, libraries must strictly comply with relevant laws and regulations such as the Personal Information Protection Law and the Data Security Law, adopt technical measures like data encryption and access control to protect user privacy, and provide privacy protection training to relevant personnel to enhance awareness and operational norms of data protection.

Furthermore, intellectual property rights and academic ethics risks are also key issues that need attention in AIGC applications. Content generated by GenAI may involve intellectual property rights issues such as copyright and patent

rights. Currently, Chinese law has not yet clearly defined the copyright ownership of AI-generated content (Li Tao, 2024). Additionally, AIGC-generated content may sometimes unintentionally infringe on others' intellectual property rights. When using these technologies, libraries must ensure that used resources are legally authorized, follow relevant intellectual property laws and regulations, and prevent copyright infringement. Meanwhile, when handling academic research and paper writing, libraries must also ensure that generated data complies with academic integrity standards to avoid academic misconduct.

The complexity of these ethical risk issues makes them difficult to completely solve through existing technical means alone, often requiring manual review, which is not only time-consuming and labor-intensive but also inefficient when facing large-scale data. Therefore, seeking efficient technical solutions has become a key focus area of current research.

2.2.2 Current Status of Ethical Security Governance Research on AIGC in Libraries

Based on the above risk considerations, some scholars have conducted technical solution research on the ethical security risks of AIGC in libraries to improve governance efficiency. However, these technical solutions are currently mainly approached from the perspective of general artificial intelligence ethical security governance frameworks. Research on AIGC ethical security governance in the library field mainly focuses on two aspects: First, preliminary governance ideas for library AIGC ethical security risks. Relevant scholars have proposed various governance strategies to address the ethical security risks of AIGC in libraries. For example, drawing on existing AI governance policies to formulate new policy recommendations (Zhang Weidong, 2023), proposing ethical principles and practical guidance suggestions (Deng Shengli, 2023), and providing AI literacy education and training (Kaushal, 2022). Additionally, supervising AIGC services through inclusive, prudent, and classified graded regulatory measures (Luo Fei, 2023), ensuring the safety and reliability of AI algorithm models and the legality of training data (Qian Yan, 2023; Li Yingting, 2023), and establishing evaluation procedures and traceable, accountable mechanisms (Lu Kang, 2021). Moreover, countries and organizations worldwide are also committed to researching general AIGC ethical security governance principles, legislation, and policies. Over 40 AI ethical principles have been proposed globally, focusing on transparency, justice, fairness, sustainability, reliability, and privacy protection. The European Artificial Intelligence Act (2024) and U.S. AI regulatory principles (2023) have provided frameworks for AI regulation and supervision. China also issued the Interim Measures for the Management of Generative Artificial Intelligence Services in 2023, specifying the responsibilities and obligations of service providers. However, these policies and regulations have not yet penetrated deeply into the library field. Furthermore, at the practical level, some libraries have formulated usage rules and norms for ChatGPT and other GenAI applications to ensure compliant use of GenAI in smart libraries. Sichuan Uni-

versity Library designed a GenAI thematic webpage, providing GenAI tools for faculty and students, while formulating usage principles and guidelines for GenAI tools from five perspectives: laws and regulations and policy guidelines, academic ethics, academic integrity, academic norms, and copyright issues[10]. In 2023, among the ChatGPT usage regulation texts issued by British and American universities, nine texts were dominated by British and American university libraries for ChatGPT policy release (Chu Jiewang et al., 2024). Most British and American university and university library websites aggregate all ChatGPT-related resources into a single module, providing source links to various related resources and explaining and regulating the use of GenAI in libraries.

Although current library AIGC ethical security governance has formed a certain systematic and practical framework, it still largely relies on general AIGC ethical technical frameworks overall. Existing AIGC ethical security technical frameworks aim to ensure AI consistency with human ethical standards through "feedback," using Reinforcement Learning (RL) methods to adjust agent behavior through Reinforcement Learning with Human Feedback (RLHF) (Ouyang, 2022). Although the RLHF method has achieved certain results, its adaptability and flexibility remain limited and cannot autonomously adjust output in real-time. Researchers have further proposed self-supervised training through AI feedback models (Bai, 2022) and ethical learning through social sandbox simulation of human social interaction feedback (Liu, 2023). However, these methods still have the problem of amplified bias in AI feedback models in practical applications, leading to the accumulation of ethical learning errors. Therefore, future research needs to further explore more adaptable and effective ethical security governance solutions.

3 Theoretical Framework for Ethical and Security Assurance

3.1 Framework Overview

To address the above ethical risks, this paper proposes an ethical security assurance theoretical framework. The core of this framework is a multi-agent collaborative governance model, which is mainly applied in smart library scenarios to solve the ethical challenges that generative AI may face in functions such as information retrieval, user consultation, and personalized recommendations. In this scenario, users first pose questions to the smart library's GenAI, and these questions, as raw input, may contain potential ethical risks. Simultaneously, the multi-agent collaborative governance model receives this input and generates supplementary prompts after multiple rounds of feedback updates. These prompts, together with the original input, are passed to the GenAI large model to guide it to generate more ethically compliant responses. The GenAI large model then outputs preliminary answers, and agents in the ethical governance technical model further conduct ethical compliance assessments of this output while deciding whether to continue supplementing prompts to optimize the out-

put based on the assessment results. After this series of processes, a final answer that meets user needs and complies with ethical standards is generated and fed back to users. This framework effectively reduces the ethical risks that GenAI may generate in smart libraries by setting up an ethical governance layer between user input and AI output, while ensuring the quality and appropriateness of responses.

The smart library AIGC ethical security assurance theoretical framework is mainly divided into two phases: the preparation phase and the operation phase. The preparation phase of the ethical security assurance theoretical framework mainly involves preparations for the multi-agent collaborative governance model, including neural network model preparation and dataset preparation. The operation phase of the ethical security assurance theoretical framework includes four main modules: ethical review, correction feedback, iterative update, and diagnostic suggestion. Each module uses different agents to work collaboratively, generating supplementary prompts for library user input content and reviewing and evaluating AIGC. Through multi-round iterative updates and feedback, the output is continuously corrected to ensure that AIGC complies with ethical standards.

3.2 Multi-Agent Collaborative Governance Model

In the current field of AIGC ethical security assurance, fine-tuning, as one of the main methods, can enable large models to follow human instructions and ethical principles (Ouyang, 2022), but it has significant drawbacks such as high costs and limited generalizability. On one hand, fine-tuning and its subsequent maintenance require substantial computational resources and rely on annotated data in specific scenarios. On the other hand, independently fine-tuned models struggle to adapt to new scenarios or tasks, meaning that large models in different scenarios must undergo specific fine-tuning to be safely applied. To address these challenges, Multi-Agent Reinforcement Learning (MARL) has gradually attracted widespread attention due to its excellent capabilities in handling complex tasks (Feng, 2024). Some scholars have begun exploring multi-agent approaches to provide more effective solutions for AIGC ethical governance. Multi-Agent Reinforcement Learning (MARL) has attracted increasing attention due to its outstanding performance in handling complex tasks and promoting collaboration among agents. In the MARL framework, multiple agents exchange information and collaborate through their respective local observations to achieve common goals. Inspired by multi-agent reinforcement learning, this study constructs an ethical security assurance framework where multi-agents collaborate from the perspective of prompts to guide GenAI models, aiming to improve generalizability across different functional scenarios. Collaboration among agents not only enhances the framework's flexibility but also enables dynamic adjustment and automatic error correction in practical applications, ensuring that output content always complies with ethical standards. As shown in Figure 1 [Figure 1: see original paper], the multi-agents proposed in this

study include detection agents, correction agents, and diagnostic agents, which work collaboratively to ensure the accuracy of model output and compliance with ethical and moral standards.

3.3 Preparation Phase

The multi-agent collaborative governance model proposed in this study mainly uses the Inverse Reinforcement Learning (IRL) algorithm (Ziebart et al., 2008; Finn et al., 2016; Zeng et al., 2022; Skalse et al., 2023) to relearn a reward function Rexpert that reflects expert values and behavioral standards, and combines both as the final reward R given to agents by the environment. The core principle of inverse reinforcement learning is to use the maximum entropy principle to infer the reward function that most likely explains expert behavior. That is, such algorithms first define a probability distribution (the reward function) covering all possible strategies, and then adjust this distribution so that actual expert behavior has the highest probability under this distribution. Considering the stability and efficiency of the algorithm in large-scale state-action spaces, this study adopts the Deep Inverse Reinforcement Learning algorithm (Guided Cost Learning, GCL) proposed by Finn et al. (2016), and the algorithm flow is shown in Table 1 .

In the preparation phase, the algorithm needs to construct two neural network models: a reward model and a policy model. The reward model takes state s and action a as input and outputs a scalar r; the policy model takes state s as input and outputs a probability distribution over the action space. Obviously, by applying gradient ascent to the reward model to maximize the likelihood function of expert trajectories, the goal of inverse reinforcement learning can be achieved. However, the algorithm still needs to explicitly construct a policy model because the optimization of the reward model involves the calculation of posterior state distributions, which depends on the unknown state transition equations and policy functions in the Markov Decision Process, making it impossible to directly calculate this state distribution. Based on the above problem, the GCL algorithm uses Monte Carlo methods to estimate this state distribution, that is, by constructing a policy function to generate a series of trajectories to estimate the state distribution. However, since the data in the dataset is generated by the current policy function, while the ideal state distribution should be driven by the current reward model, the difference between the two may significantly affect the accuracy of state distribution estimation. To solve this problem, the GCL algorithm introduces importance sampling technology to adjust the estimated values to make them closer to the actual state distribution driven by the reward model.

In addition, the algorithm needs to maintain three datasets: expert dataset Ddemo, non-expert dataset Dsamp, and policy dataset Dtraj. The expert dataset is a pre-collected collection of examples where functional large models output ethically compliant responses through the addition of supplementary prompts. The policy dataset consists of trajectories generated by agents ac-

cording to the current policy model, reflecting the agents' immediate behavior patterns. The non-expert dataset includes the aggregation of all policy datasets to date, recording the behavior changes of agents under different policy iterations. During the optimization phase of the GCL algorithm, agents first generate trajectories according to the current policy to form the policy dataset Dtraj and merge this dataset into the non-expert dataset. Then, the reward model is optimized for k steps. Specifically, step one involves sampling a trajectory collection from the non-expert dataset and the expert dataset respectively to form collection D, and step two involves calculating gradients based on the trajectories in this collection to update the reward model parameters. These k updates aim to adjust the reward model parameters so that the return of trajectories generated based on the current reward model is higher than that of non-expert trajectories, approaching or exceeding expert trajectories as much as possible. Next is the policy model update, which uses a dataset entirely sampled from the policy dataset and adopts a model-free reinforcement learning method. This update helps improve policy performance, ensures that generated trajectories are more optimized, and makes agent behavior more efficient and goal-oriented. The above steps are executed cyclically until preset performance indicators are reached or the algorithm converges.

By combining binary rewards and rewards learned through inverse reinforcement learning, correction agents can not only respond based on static preset rules but also dynamically adapt to new environmental conditions and moral considerations, significantly improving the reliability and ethical compliance of the model in practical applications. After the new reward function is defined, the reinforcement learning algorithm PPO (Proximal Policy Optimization) (Schulman et al., 2017), which has high stability and efficiency, is used to continue training correction agents based on the policy network obtained from inverse reinforcement learning training.

3.4.1 Ethical Review Module

The ethical review module is responsible for automatically monitoring the output of generative AI models by detection agents. Its main task is to evaluate the ethical compliance of output content. This agent learns from an annotated expert behavior dataset to build a binary classification model that can analyze new large model outputs in real-time. Specifically, the detection agent compares the input text with texts annotated as "compliant" or "non-compliant" with ethical standards in its training set to determine the compliance of output content. During the process of library users asking questions and receiving feedback from GenAI large models, when generated content complies with ethical standards, the detection agent will directly provide it to users, ensuring that the content they receive is safe and compliant. However, if the output content does not meet ethical standards, the detection agent will feed this information back to the correction agent for processing. In this process, the detection agent continuously updates its model parameters through an ongoing online learn-

ing mechanism to adapt to newly emerging ethical standards and user needs. This dynamic adjustment capability enables the detection agent to effectively respond to constantly changing ethical requirements, enhancing its judgment ability and response speed.

3.4.2 Correction Feedback Module

The correction feedback module is mainly completed by correction agents, whose core task is to receive feedback from detection agents and generate supplementary prompts based on this feedback to guide GenAI large models to produce ethically compliant output. Correction agents combine the Inverse Reinforcement Learning (IRL) algorithm to create a reward function (Rexpert) that reflects expert values. This reward function is combined with the original binary reward (Rbinary) to form the final reward (R) to guide the agent's learning process. The working mechanism of correction agents can be formalized as a Markov Decision Process (MDP), where the state is jointly constituted by the user's initial query and the GenAI large model's response. The agent's action is to generate supplementary prompts aimed at guiding the large model to output ethically compliant content. Since this agent faces the problem of reward sparsity, the inverse reinforcement learning algorithm (GCL) is introduced to optimize the training process. The GCL algorithm constructs two neural network models—a reward model and a policy model—to maximize the likelihood of expert behavior. The reward model takes state (s) and action (a) as input and outputs a scalar (r), while the policy model takes state (s) as input and outputs a probability distribution over the action space. By applying gradient ascent to the reward model to optimize expert trajectories, the agent can effectively learn supplementary prompt generation strategies that comply with ethical standards during training.

3.4.3 Iterative Update Module

The iterative update module is the core of the entire security assurance technical model, aiming to ensure that the output of diagnostic agents meets ethical requirements through dynamic feedback and continuous error correction. This module combines feedback from detection agents and correction agents to continuously update the reward function (R) of correction agents, enabling them to adapt to new environmental conditions and ethical standards. In practice, the iterative update module continuously collects user feedback and outputs from generative large models, analyzes which content fails to meet ethical standards, and adjusts the strategies of correction agents accordingly. The module also adopts model-based reinforcement learning methods to ensure that correction agents can obtain new learning opportunities in each iteration, thereby improving their adaptability in complex and dynamic environments. Through this approach, the iterative update module not only improves model reliability but also promotes overall system optimization and perfection, ensuring that output content always complies with ethical standards. Ultimately, the goal

of this module is to build an adaptive ethical compliance mechanism that can effectively respond to potential ethical challenges in the future.

3.4.4 Diagnostic Suggestion Module

The diagnostic suggestion module is responsible for diagnostic agents, whose main function is to improve the accuracy of predicting ethical issues that may be caused by user questions by minimizing cross-entropy loss. Diagnostic agents are trained using a large amount of annotated data, including various user questions and related model responses that have been evaluated by human experts and annotated as to whether they involve ethical issues and specific types (such as cultural insensitivity, academic misconduct, or intellectual property infringement). During the training process, diagnostic agents continuously optimize their predictive ability through end-to-end supervised learning methods. After users submit questions, diagnostic agents analyze and predict potential ethical risks in real-time and provide feedback and guidance to users. This process not only focuses on the compliance of model output but also guides users to understand the ethical issues that may be triggered by their questions. If, after multiple iterations, the output of the generative large model still poses ethical risks, the diagnostic agent will proactively alert users to potential ethical issues. For example, the agent may suggest that users change their questioning method or provide more appropriate expression methods to reduce ethical risks. Through this approach, diagnostic agents not only improve users' ethical awareness but also promote a healthier interactive environment. Additionally, the diagnostic suggestion module will collect and analyze user questions, functional large model responses, and text data generated by agents. This data will be used for subsequent supervision, management, and research analysis to continuously optimize the system's ability to follow ethical standards, ensuring that the model's ethical compliance is continuously improved in future applications.

4 Conclusion

In the field of smart libraries, the application of Generative Artificial Intelligence (GenAI) has significantly driven service innovation and efficiency. However, the widespread application of this technology has also triggered a series of ethical and security issues, urgently requiring the establishment of a systematic assurance mechanism. This paper constructs an ethical security assurance theoretical framework based on multi-agent collaborative governance, providing solutions for GenAI applications in smart libraries. In this study, we first reviewed the current application status of GenAI in smart libraries and found that although this technology has significant advantages in service innovation, knowledge management, and user experience enhancement, it also faces ethical security challenges such as output authenticity, privacy protection, intellectual property rights, and academic ethics. To address these issues, we proposed an ethical security assurance framework based on a multi-agent collaborative governance model. The research results of this paper indicate that current research on ethical security

assurance of GenAI in smart libraries mainly focuses on the theoretical level and has not yet formed a unified solution. The proposed theoretical framework provides a systematic solution for the ethical security management of GenAI in smart libraries, helping to ensure the ethically compliant application of GenAI technology and thus promoting the long-term development of smart libraries. By implementing this framework, manual review workload can be effectively reduced, processing efficiency improved, and the accuracy and ethical compliance of generated content ensured. Nevertheless, this study still has some limitations. The existing framework has not yet been extensively validated in practical applications, and its effectiveness and applicability need to be verified through practice in subsequent research.

References

[1] An Zidong, Jing Qing, Hao Zhichao, et al. Innovative Strategies for Library Literature Resource Management Based on Generative AI Technology[J]. Library Work and Study, 2023, (S1): 9-16.

[2] Chu Jiewang, Du Xiuxiu. Enlightenment of ChatGPT Guidance in British and American Universities to Chinese University Libraries[J]. Library and Information Service, 2024, 68(5): 42-53.

[3] Deng Shengli, Wang Fan. Research Progress and Development Trends of AIGC Governance[J]. Digital Library Forum, 2023, 19(11): 20-28.

[4] Gong Furong. Exploring the Impact of ChatGPT-like Generative AI on Digital Literacy Education in University Libraries[J]. Document, Information & Knowledge, 2023, 40(5): 97-106+156.

[5] Guo Yajun, Guo Yiruo, Li Shuai, et al. ChatGPT Empowering Smart Library Services: Characteristics, Scenarios, and Paths[J]. Library Development, 2023, (2): 30-39+78.

[6] Guo Yajun, Kou Xuying, Feng Siqian, et al. Large Language Model Empowers Library Reference Services: Logic, Scenarios, and Framework[J]. Library Tribune, 2024, 1-10.

[7] Hou Zhijiang. Application Modes and Implementation Paths of ChatGPT in Library[J]. Library Theory and Practice, 2024, (3): 102-110+127.

[8] Li Shuning, Liu Yiming. Opportunities and Challenges for Library from the Rise of ChatGPT-like Intelligent Chat Tools[J]. Library Tribune, 2023, 43(5): 104-110.

[9] Li Tao. The Opportunity, Risk and Legal Regulation of Generative Artificial Intelligence Sora to Intelligent Library[J]. Library Development, 2024(7): 1-10.

[10] Li Yingting. Opportunities, Challenges, and Strategies for Libraries in the Era of Artificial Intelligence Generated Content[J]. Library & Information, 2023, (2): 42-48.

[11] Liu Li, Shao Bo. Exploring the Convergence Path between Generative AI and Smart Libraries — Taking Zayed University Library as an Example[J]. Researches In Library Science, 2023, (12): 34-43.

[12] Lu Kang, Liu Hui, Zhang Jing, et al. Internal Path Analysis of Library Smart Service Based on AI Governance: Taking the EU 'Artificial Intelligence Ethics Code' as an Example[J]. Library Theory and Practice, 2021, (6): 55-60.

[13] Luo Fei, Cui Bin, Xin Xiaojiang, et al. The Risk Paradigm and Management Strategies of Embedding Big Language Model into Library Knowledge Services[J]. Library & Information, 2023, (3): 99-106.

[14] Qian Yan, Mei Ying. From Concept to Practice: Exploring the Application of Generative Artificial Intelligence in Smart Libraries[J]. Library Science Research & Work, 2023, (12): 27-34.

[15] Sun Shangfeng, Shen Ning, Liu Lin, et al. Research on the Promotion Model of Library Intelligent Reading in the Age of Generative AI[J]. Library Theory and Practice, 2024, (4): 83-88.

[16] Wang Jing, Wang Peng. Research on Risk Management Mechanism of Generative AI Large Model in Smart Library[J]. Journal of Intelligence, 1-8.

[17] Wu Ruohang, Mao Yihong. Library Services under the ChatGPT Boom: Concepts, Opportunities, and Breakthroughs[J]. Library & Information, 2023, (2): 34-41.

[18] Xu Fang. Application Scenarios and Legal Issues of Generative Artificial Intelligence in Smart Libraries[J]. Information and Documentation Services, 2024, 45(2): 24-29.

[19] Zhang Hai, Duan Hui, Wang Dongbo. Artificial Intelligence Generative Content in Information Resource Management: Research Status, Research Topics, and Research Prospect[J]. Journal of Intelligence, 2024(7): 1-7.

[20] Zhang Hui, Tong Tong, Ye Ying. GPT-Driven Technical Innovation of Smart Libraries in the AI 2.0 Era[J]. Library Journal, 2023, 42(5): 4-8.

[21] Zhang Weidong, Chen Xipeng, Zhao Hongying. Global Library Science in the Era of Data Intelligence: Current Status, Hotspots, and Trends– The Reviews of 88th IFLA World Library and Information Congress 2023[J]. Information Science, 2023, 41(12): 1-8+32.

[22] Zhou Xu. Opportunities and Challenges: Analysis on Libraries' Response under the Background of ChatGPT Popularization[J]. Library, 2023, (6): 34-41+48.

[23] Zhu Yu, Chen Guanze, Lu Yongrong, et al. Generative Artificial Intelligence Governance Action Framework: Content Analysis Based on AIGC Incident Report Texts[J]. Document, Information & Knowledge, 2023, 40(4): 41-51.

[24] Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: Harmlessness from AI feedback[J]. arXiv preprint, arXiv: 2212.08073, 2022.

[25] Feng P, Liang J, Wang S, et al. Hierarchical Consensus-Based Multi-Agent Reinforcement Learning for Multi-Robot Cooperation Tasks. ArXiv. 2024.

[26] Finn C, Levine S, Abbeel P. Guided cost learning: Deep inverse optimal control via policy optimization[C]//International conference on machine learning. PMLR, 2016: 49-58.

[27] Kaushal V, Yadav R. The Role of Chatbots in Academic Libraries: An Experience-based Perspective[J/OL]. Journal of the Australian Library and Information Association, 2022, 71(3).

[28] Liu R, Yang R, Jia C, et al. Training Socially Aligned Language Models on Simulated Social Interactions[C]//The Twelfth International Conference on Learning Representations. 2023.

[29] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in neural information processing systems, 2022, 35: 27730-27744.

[30] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.

[31] Skalse J, Abate A. Misspecification in inverse reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(12): 15136-15143.

[32] Yrjo Lappalainen, Narayanan N. Aisha: A Custom AI Library Chatbot Using the ChatGPT API[J]. Journal of Web Librarianship, 2023, 17(3): 1-22.

[33] Zeng S, Li C, Garcia A, et al. Maximum-likelihood inverse reinforcement learning with finite-time guarantees[J]. Advances in Neural Information Processing Systems, 2022, 35.

[34] Ziebart B D, Maas A L, Bagnell J A, et al. Maximum entropy inverse reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2008, 8.

---

[1] https://genai.umich.edu/

[2] https://news.harvard.edu/gazette/story/newsplus/harvard-designs-ai-sandbox-that-enables-exploration-interaction-without-compromising-security/

[3] https://www.thecrimson.com/article/2023/9/1/fas-ai-guidance/

[4] https://libguides.princeton.edu/c.php?g=1341922&p=10191899

[5] https://zhuanlan.zhihu.com/p/422906714

[6] https://lib.ccnu.edu.cn/info/1072/7542.htm

[7] http://www.bhwh.gov.cn/home/content/detail/id/36775.html

[8] https://chatlibrary.newacademic.net

[9] Science and Technology. Risks and Mitigation Strategies for Adversarial

Artificial Intelligence Threats: A DHS S&T Study. 2023.12.12.
[10] https://lib.scu.edu.cn/genai/care.html

Note: Figure translations are in progress. See original paper for figures.