


Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information

Glorin Sebastian, Georgia Institute of Technology, USA*

 <https://orcid.org/0000-0003-2543-9127>

ABSTRACT

The evolution of artificial intelligence (AI) and machine learning (ML) has led to the development of sophisticated large language models (LLMs) that are used extensively in applications such as chatbots. This research investigates the critical issues of data protection and privacy enhancement in the context of LLM-based chatbots, with a focus on OpenAI's ChatGPT. It explores the dual challenges of safeguarding sensitive user information while ensuring the efficiency of machine learning models. It assesses existing privacy-enhancing technologies (PETs) and proposes innovative methods, such as differential privacy, federated learning, and data minimization techniques. The study also includes a survey of Chatbot users to measure their concerns related to data privacy with the use of these LLM-based applications. This study is meant to serve as a comprehensive guide for developers, policymakers, and researchers, contributing to the discourse on data protection in artificial intelligence.

KEYWORDS

Artificial Intelligence, ChatGPT, Cybersecurity, Data Protection, Large Language Model

1. INTRODUCTION

“ChatGPT, developed by OpenAI in November 2022, is an AI chatbot that utilizes the Generative Pre-trained Transformer (GPT) model. OpenAI is an AI research and development company known for its innovative approaches in natural language processing. The GPT model, based on the Transformer architecture introduced by Vaswani et al. (2017), is trained on extensive datasets to generate contextually relevant and accurate responses to text-based inputs. However, as these systems become more sophisticated and widely used, concerns regarding user privacy and data protection have emerged. Large Language Models (LLMs) like ChatGPT aim to understand and generate human language, but their reliance on extensive datasets, which may contain sensitive information, raises privacy concerns. There is a risk of inadvertently capturing and exposing sensitive user data, particularly in the context of chatbots and virtual assistants where personal or confidential information is often disclosed. These concerns have been addressed in various research papers discussing the usage of LLM-based chatbots, such as those by Hariri (2023), Sebastian (2023), and Cao et al. (2023).

DOI: 10.4018/IJSPPC.325475

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

However, these research papers have not delved deeply into the topic of data privacy risks in LLM chatbots, this paper addresses this research gap by reviewing the data privacy risks associated with LLM chatbots. Further, to mitigate these privacy concerns, it is essential to develop effective strategies and technologies that can safeguard user data while maintaining the utility of LLMs. This paper aims to address this need by examining current privacy concerns, exploring existing privacy-enhancing technologies (PETs), and proposing novel techniques to ensure robust data protection in LLMs like ChatGPT. The techniques include differential privacy, federated learning, data minimization, and secure multi-party computation. Additionally, this research explores legal and ethical frameworks that can guide the responsible development of AI systems, considering both the tremendous potential of LLMs and the importance of user privacy. The paper serves as a comprehensive guide for developers, policymakers, and researchers in this rapidly evolving field, contributing to the ongoing dialogue about data protection in AI and promoting the development of innovative technologies that prioritize user privacy.”

1.1 Brief Overview of ChatGPT

ChatGPT is an advanced AI model developed by OpenAI, which utilizes the Generative Pretrained Transformer (GPT) series of models. GPT belongs to the category of large language models (LLMs), which are characterized by their extensive training on diverse and comprehensive linguistic datasets and their ability to generate human-like text that is contextually relevant and coherent. ChatGPT, specifically, is designed to engage in conversation with users, with applications ranging from virtual assistants and customer service bots to AI tutors and more. The power of ChatGPT lies in its capacity to understand and generate meaningful responses to a wide array of prompts, demonstrating a deep grasp of syntax, semantics, and even nuanced aspects of conversation such as humor and emotion. The ChatGPT is a closed model without information about its training dataset and how it is currently being trained. Preventing data leakage (training-test contamination) is one of the most fundamental principles of Machine learning because such leakage makes evaluation results unreliable (Aiyappa, Rachith, et al.,2023).

Training an LLM like ChatGPT involves two main steps: pre-training and fine-tuning (Zheng, Ou, et al.,2023). During pre-training, the model is exposed to a large corpus of Internet text to learn grammar, facts about the world, reasoning abilities, and unfortunately, some of the biases present in the training data. In the fine-tuning process, ChatGPT is further trained on a narrower dataset, generated with the help of human reviewers following specific guidelines provided by OpenAI. Despite its impressive capabilities, ChatGPT, like all AI systems, raises some important privacy and data protection issues. Since the model learns from vast amounts of data, there is a risk of it inadvertently learning and generating sensitive or personally identifiable information. Also, user interactions with ChatGPT could potentially expose personal data, either through the questions users ask or the context in which the system is deployed. Hence, it is critical to explore techniques and strategies to enhance privacy and data protection in ChatGPT and similar LLMs.

1.2 Importance of Privacy and Data Protection in AI Systems

The significance of privacy and data protection in AI systems cannot be overstated and encompasses ethical, legal, and user trust considerations. Preserving privacy and data protection is crucial for the ethical, responsible, and compliant development and deployment of AI systems, contributing to user trust and the overall success and acceptance of these technologies.

- i) **Ethical Considerations:** Ethical principles dictate that the personal and sensitive data of users should be respected and safeguarded. AI systems, especially large language models (LLMs) that process vast amounts of data, may unintentionally learn and generate sensitive information. Any compromise of personal data can lead to detrimental consequences such as identity theft,

12 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/privacy-and-data-protection-in-chatgpt-and-other-ai-chatbots/325475

Related Content

Comparing Two Playability Heuristic Sets with Expert Review Method: A Case Study of Mobile Game Evaluation

Janne Paavilainen, Hannu Korhonen and Hannamari Saarenpää (2012). *Media in the Ubiquitous Era: Ambient, Social and Gaming Media* (pp. 29-52).

www.irma-international.org/chapter/comparing-two-playability-heuristic-sets/58579

Introduction to Smart Phone Positioning

Ruizhi Chen (2012). *Ubiquitous Positioning and Mobile Location-Based Services in Smart Phones* (pp. 1-31).

www.irma-international.org/chapter/introduction-smart-phone-positioning/67037

U-Learning Pedagogical Management: Cognitive Processes and Hypermediatic Resources Involved in Web-Based Collaborative Workspace

Jocelma Almeida Rios, Emanuel do Rosário Santos Nonato, Mary Valda Souza Sales and Tereza Kelly Gomes Carneiro (2014). *Technology Platform Innovations and Forthcoming Trends in Ubiquitous Learning* (pp. 270-288).

www.irma-international.org/chapter/u-learning-pedagogical-management/92948

Warranting High Perceived Quality of Experience (PQoE) in Pervasive Interactive Multimedia Systems

Anxo Cereijo Roibás (2010). *Ubiquitous and Pervasive Computing: Concepts, Methodologies, Tools, and Applications* (pp. 1498-1516).

www.irma-international.org/chapter/warranting-high-perceived-quality-experience/37864

Monitoring and Optimization of Pilot Pollution in High-Rise

Tianze Li, Tao Gao, Ye Liu, Yuhang Wang and Jiahui Chen (2016). *International Journal of Advanced Pervasive and Ubiquitous Computing* (pp. 87-128).

www.irma-international.org/article/monitoring-and-optimization-of-pilot-pollution-in-high-rise/176605