A Framework for Integrating Privacy by Design into Generative AI Applications

Hamda Al Breiki[1], Qusay H. Mahmoud[2]

[1]Zayed University
[2]Ontario Tech University
hamda.albreiki@zu.ac.ae, qusay.mahmoud@ontariotechu.ca

Abstract

Generative AI applications interact with user-provided data, raising privacy concerns due to potential exposure of sensitive information. Traditional privacy safeguards often follow a reactive approach, addressing risks only after deployment. However, given the evolving nature of AI-driven data processing, a proactive and systematic approach to privacy integration is necessary. This paper presents a framework for embedding principles of Privacy by Design (PbD) and other privacy mechanisms throughout the AI lifecycle. Unlike traditional PbD implementations that primarily focus on data collection and storage, the proposed framework introduces privacy-preserving techniques at the model level, ensuring AI models minimize data exposure during training and inference. We propose dynamic user consent mechanisms, differential privacy-enhanced model architectures, federated learning for decentralized training, and real-time privacy risk monitoring tools to enhance transparency, security, and user control. Additionally, the framework incorporates fairness-aware privacy techniques, ensuring that privacy measures do not exacerbate bias in AI models. The framework is evaluated through empirical testing of privacy leakage risks and differential privacy tradeoff analysis. Results demonstrate that integrating PbD like mechanisms into generative AI enhances privacy protections while maintaining AI utility and regulatory compliance.

Introduction

A generative AI application is a software solution that utilizes generative models to produce new content (text, images, audio, or even video) by learning patterns from training data. These applications can range from creative tools, like art or music generators, to practical systems such as chatbots, virtual assistants, or data augmentation tools. They often leverage architectures like transformers, Generative Adversarial Networks, or Variational Autoencoders (VAEs) to generate outputs that are contextually relevant and sometimes entirely novel (Sengar et al., 2024). The adoption of GenAI applications has accelerated across industries such as healthcare, education, finance, and entertainment, enabling advancements in personalized services, decision-making, and creative innovation. Notable examples, including ChatGPT, DALL-E, and Midjourney, illustrate the profound impact of generative AI on automation and human-AI collaboration. However, despite their potential, these systems raise significant privacy concerns due to their reliance on vast datasets, often containing personal or sensitive information, to train and refine models (Park et al., 2025).

The increasing frequency of data breaches, unauthorized disclosures, and privacy violations highlights the urgent need for robust privacy safeguards in generative AI. In high-risk sectors such as healthcare and finance, where regulatory frameworks like the General Data Protection Regulation (GDPR, 2016) and the UAE Personal Data Protection Law (PDPL, 2021) impose strict privacy requirements, compliance challenges become particularly pronounced. Existing privacy-preserving strategies remain fragmented, often addressing privacy concerns only after deployment, rather than being embedded into AI systems from the outset.

To address these challenges, this paper presents a privacy framework tailored for generative AI, ensuring privacy is proactively integrated into AI lifecycle. Unlike conventional privacy safeguards that focus solely on data collection and access control, this framework extends privacy protections to the AI model level, incorporating differential privacy, federated learning, and dynamic consent management. By embedding privacy mechanisms into the entire AI lifecycle, the framework enhances transparency, accountability, and user trust while ensuring compliance with global regulations.

The proposed framework is demonstrated through a privacy-aware generative AI chatbot that integrates key PbD principles to enhance user control and privacy protections. The chatbot offers tiered privacy modes: strict, standard, and personalized, ensuring data retention and processing align with individual preferences. Differential privacy mechanisms prevent the AI from memorizing personal data, reduc-

ing sensitive information leakage, while privacy risk detection tools warn users when they input Personally Identifiable Information (PII), and encrypted logging maintains regulatory compliance. This implementation shows how generative AI systems can incorporate privacy-preserving architectures without compromising functionality or user experience. The evaluations highlight improvements in privacy compliance, transparency, and accountability, offering a scalable approach for responsible AI development.

Contributions

The primary novelty of this paper lies in the proactive and systematic integration of PbD principles into generative AI workflows at multiple levels (data, model, and inference stages). It moves beyond traditional PbD approaches, which focus on data governance or post-hoc regulatory compliance. The contributions can be summarized as follows:

- *Embedding privacy at the model level*: Integrating differential privacy, federated learning, and secure multiparty computation within the AI architecture itself rather than treating privacy as an external, data-level concern.
- *Dynamic consent management*: Offering tiered user consent settings (strict, standard, personalized) allowing dynamic adjustment based on context and user preference.
- *Real-time privacy monitoring*: Employing automated tools and continuous risk monitoring rather than post-deployment audits.
- *Fairness-aware privacy*: Addressing the often-overlooked privacy-fairness tradeoff to avoid disproportionate bias against specific demographic groups.

Privacy Implications of AI-Generated Content

Generative AI systems present unique privacy challenges that extend beyond data collection and training, reaching into the outputs they generate. AI-produced content can inadvertently reveal private or sensitive information. Large language models (LLMs) have been shown to memorize parts of their training data, occasionally PII such as email addresses, social security numbers, or private messages when confronted with adversarial prompts. This unintentional exposure not only raises ethical concerns but also poses significant regulatory challenges under frameworks like the GDPR, which mandates the "Right to Be Forgotten." High-profile cases such as the GitHub Copilot controversy, where sensitive information, including private API keys and proprietary code, was unintentionally reproduced, highlight the need for robust privacy monitoring and automated content filtering mechanisms in generative AI systems.

Moreover, the advent of sophisticated generative adversarial networks (GANs) and diffusion models has led to the creation of highly realistic deepfakes, which amplify privacy risks by enabling the unauthorized replication of personal attributes such as likeness, voice, and style. These synthetic media forms can facilitate identity theft, defamation, and fraud by allowing malicious actors to impersonate real individuals without consent. The legal and ethical gaps surrounding AI-generated content further complicate privacy protections, as many jurisdictions lack comprehensive regulations addressing deepfake technologies. Together, these issues highlight the urgent need for privacy-preserving strategies and policy interventions that ensure responsible deployment of generative AI systems, balancing innovation with robust privacy safeguards.

Background and Related Work

Privacy by Design (PbD) has become increasingly recognized as a fundamental principle for developing privacy-aware AI systems, given the sensitivity and volume of data involved. This review discusses existing literature to clarify the relationship between privacy practices, ethical considerations, and generative AI.

The work by Feretz et al. (2024) highlights the necessity for privacy models tailored explicitly to generative AI contexts, recognizing that different phases of the AI lifecycle demand distinct privacy-preserving techniques. This perspective complements insights from Huriye (2023), who emphasizes transparency and accountability as foundational elements for ethically deploying AI, ensuring consistent responsiveness to user privacy concerns across system development phases.

Sripras et al. (2024) extend this discussion by highlighting secure methodologies for data sharing within commercial AI applications, aligning with regulatory frameworks such as GDPR. They advocate for methods that enable secure data exchanges while preserving competitive insights, crucial for maintaining user trust and regulatory compliance. Additionally, ethical implications and potential misuse of generative AI, including risks of misinformation, are critically examined by Luk et al. (2024), reinforcing the societal imperative for robust ethical regulation.

Ijiga et al. (2024) further expand the above works by addressing ethical concerns in healthcare deployments of generative AI, identifying complexities tied to data privacy, transparency, and equitable implementation across diverse regulatory environments. Vallverdú (2023) also reinforces these concerns by examining the ethical deployment of generative AI within healthcare, highlighting specific challenges associated with sensitive data handling and emphasizing the critical role of privacy-preserving designs.

Privacy and ethical considerations are further explored by Olorunsogo et al. (2024), who argue for trust as a pillar of AI-enhanced medical decision-making systems. They highlight that embedding privacy considerations into AI frameworks enhances transparency and user confidence. This argument aligns with Xu et al. (2021), who call attention to the importance of privacy-preserving methodologies within machine learning paradigms, advocating for a balanced approach that achieves operational efficiency without compromising privacy. Yu et al. (2023) extend this narrative by proposing comprehensive frameworks that integrate privacy-preserving methods into the core design of generative AI,reflecting a growing consensus on the need for balance between operational efficiency and ethical standards. Similarly, Vallverdú (2023) illustrates the ethical challenges associated with deploying generative AI in healthcare, emphasizing the necessity of incorporating robust privacy mechanisms to mitigate associated risks effectively.

In a survey paper by Das et al. (2025), a detailed overview of LLMs' security and privacy challenges was presented. The authors found that privacy risks in LLMs arise from their inherent capacity to process and generate text based on extensive and diverse training datasets. They stated that key challenges in LLM privacy are Data memorization, data leakage, and the potential disclosure of confidential information. Additionally, the paper proposed future research directions focusing on security and privacy aspects of LLMs. Gupta et al. (2023) also explore the implications associated with the use of Generative AI in the fields of cybersecurity and privacy with a focus on ChatGPT. The paper considered privacy and data protection as one of the open research challenges for GenAI and LLMs.

Park et al. (2025) introduced the Context-Aware Privacy Framework for Multi-Agent Generative AI Applications (CAPRI). This framework integrates a local gatekeeper LLM responsible for pseudonymizing PII and sensitive data within entity structures before any interaction occurs with a cloud-based third-party LLM. By doing so, it enables LLM agents to securely handle user queries while safeguarding data privacy. Additionally, CAPRI features a private, local, and encrypted storage system that maintains records of the pseudonymized entities, supporting reversible mapping through the use of a unique key.

While Privacy by Design (PbD) has been widely explored in data governance and access control, its integration into generative AI models remains underdeveloped. Most existing frameworks focus on privacy compliance, data access restrictions, and regulatory adherence but do not embed privacy mechanisms at the model architecture level. Specifically, there is a lack of systematic approaches for integrating differential privacy, federated learning, and secure multi-party computation into generative AI workflows, leaving these models vulnerable to privacy breaches and data leaks. Another key limitation is the absence of dynamic user privacy controls. Traditional AI privacy implementations rely on static, binary consent models (opt-in/opt-out), without enabling users to adjust settings based on contextual risk factors. Research on tiered privacy modes that allow a balance between protection and utility is limited. In addition, current PbD implementations emphasize post-deployment audits rather than real-time privacy monitoring. Existing approaches lack AI-driven privacy risk assessment tools capable of detecting privacy violations during inference.

Moreover, privacy-enhancing techniques often overlook the privacy-fairness tradeoff. For instance, differential privacy may introduce algorithmic bias that disproportionately affects underrepresented groups. There is a need for fairness-aware privacy techniques that ensure robust protections while preserving model fairness. Finally, although theoretical discussions on PbD exist, practical implementation challenges remain unresolved, particularly regarding scalable integration into production-level generative AI systems without compromising efficiency, performance, or usability.

This paper addresses these gaps by introducing a framework for operationalizing PbD mechanisms tailored for generative AI. Our framework extends privacy beyond data governance to include AI model architectures, introduces dynamic user-controlled privacy settings, incorporates real-time privacy risk monitoring, and integrates fairness-aware differential privacy techniques. We demonstrate the feasibility of our approach through a privacy-aware chatbot prototype, validating its applicability in real-world scenarios.

Framework Design

This proposed framework is designed to proactively embed privacy mechanisms throughout the entire AI lifecycle. Unlike conventional PbD frameworks that focus primarily on privacy compliance and data management, the proposed framework consists of the following components (Figure 1).

- *Proactive Privacy Integration*. A fundamental principle of the framework is proactive privacy integration, ensuring that potential risks are identified and mitigated before AI systems are deployed. Traditional AI development often follows a sequential pipeline where privacy considerations emerge only in later stages, typically during compliance assessments. Our framework, by contrast, mandates early-stage privacy impact assessments to evaluate potential privacy risks associated with data collection, model training, and deployment. To minimize personal data exposure, the framework advocates for data minimization techniques, such as synthetic data

generation and privacy-preserving data preprocessing. Additionally, differential privacy mechanisms are incorporated during training, ensuring that individual data points cannot be reconstructed from model outputs. These measures collectively reduce the risk of privacy breaches while maintaining AI performance.

- *Data Transparency and Explainability*. One of the most significant barriers to privacy adoption in AI systems is the lack of transparency regarding data handling and model decisions. Generative AI models are often trained on vast datasets, yet users remain unaware of how their data is processed, stored, and repurposed. The proposed framework incorporates explainable AI (XAI) techniques to enhance transparency, making AI decisions more interpretable. To further improve transparency, the framework proposes privacy-aware data provenance tracking using blockchain-based logging. This ensures that all interactions with user data are auditable, tamper-proof, and traceable, allowing users and regulatory bodies to verify compliance. Additionally, layered privacy policies, combining high-level summaries with detailed technical explanations, enhance accessibility, ensuring users can make informed decisions about data sharing.
- *User Control and Consent Management*. Ensuring user autonomy over personal data is a core objective of the framework. Traditional AI systems often provide only binary consent options, failing to address nuanced user preferences. To resolve this, the framework introduces granular consent management mechanisms that allow users to define specific data usage permissions. The framework also incorporates real-time consent dashboards, enabling users to review, modify, or revoke consent dynamically. These dashboards provide insights into which data points are being used, the AI model's learning process, and how privacy settings impact AI performance. Furthermore, by offering tiered privacy modes (e.g., Strict Privacy, Standard, and Personalized AI), users can balance privacy concerns with AI utility based on their preferences.
- *Privacy-Preserving AI Architectures*. A significant technical challenge in privacy-conscious AI development is designing secure model architectures that mitigate data exposure risks. The proposed framework integrates multiple privacy-preserving techniques at the architectural level, including: (1) Federated Learning: Decentralizes AI training by keeping user data on local devices, thereby reducing the risk of centralized data breaches. (2) Secure Multiparty Computation (SMPC): Enables AI models to process encrypted data without decrypting it, ensuring that sensitive information remains protected. And (3) On-Device AI Processing: Shifts computations from cloud environments to user-controlled devices, minimizing third-party data access.

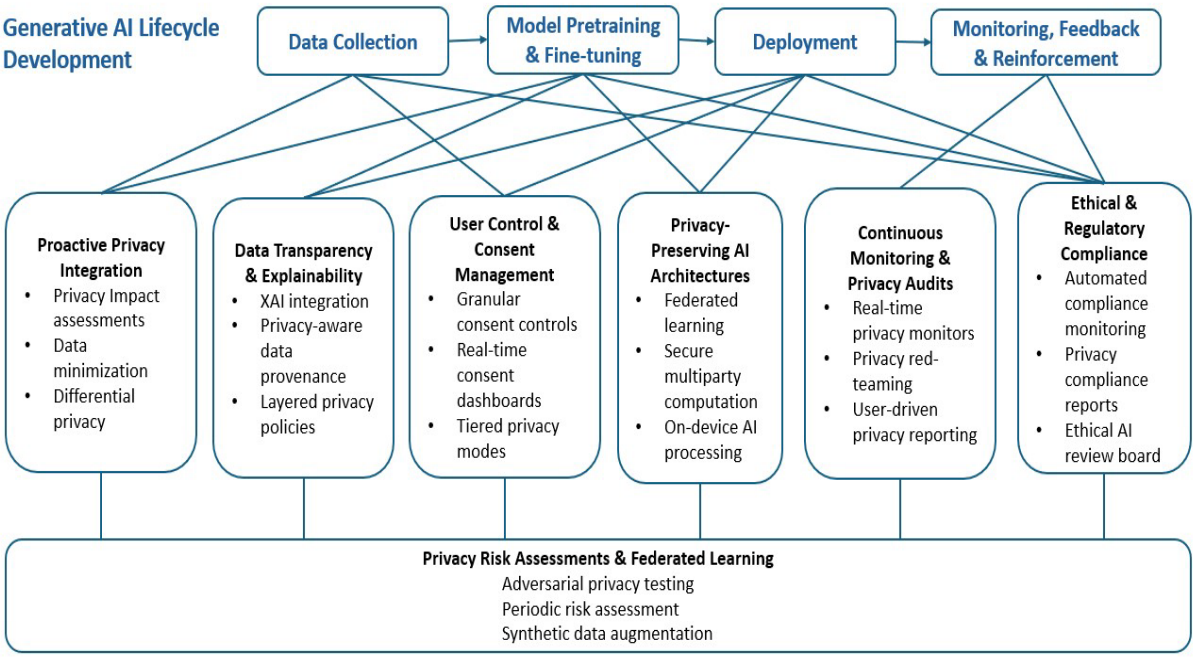

Figure 1: Components of the proposed framework.

- *Continuous Monitoring and Privacy Audits*. Given the evolving nature of AI privacy threats, our framework incorporates continuous monitoring and adaptive risk mitigation mechanisms. Instead of relying solely on periodic audits, the framework integrates automated privacy monitoring tools that analyze AI interactions in real time to detect and mitigate emerging privacy risks. To strengthen accountability, the framework introduces privacy red-teaming, a proactive security strategy where ethical hackers simulate privacy attacks to identify and patch vulnerabilities before they can be exploited. Additionally, user-driven privacy reporting mechanisms are embedded within AI interfaces, allowing individuals to report privacy concerns directly and receive prompt responses from AI system administrators.
- *Privacy, Ethical and Regulatory Compliance Alignment.* The increasing global emphasis on AI regulations necessitates compliance with privacy laws such as GDPR, the California Consumer Privacy Act (CCPA), and the UAE Personal Data Protection Law (PDPL). Our framework aligns generative AI development with these regulations by establishing automated compliance monitoring mechanisms. A key aspect of this compliance strategy is the Privacy Compliance Report (PCR), which documents how AI models adhere to privacy laws throughout their lifecycle. This report is regularly updated and auditable, providing regulators, developers, and users with clear insights into AI privacy practices. Furthermore, the framework advocates for the creation of an Ethical AI Review Board, an independent oversight body that ensures privacy and ethical standards are continuously upheld.
- *Privacy Risk Assessments and Federated Learning Adoption*. To ensure continuous privacy protection, the proposed framework mandates periodic privacy risk assessments to evaluate vulnerabilities in AI models. These assessments include adversarial stress testing, where models are exposed to privacy attacks (e.g., membership inference attacks, model inversion attacks) to determine their resilience against data leaks. Additionally, the framework promotes the adoption of federated learning not only for initial model training but also for ongoing updates, reducing reliance on centralized data repositories. To further limit data exposure, the framework encourages the use of synthetic data augmentation, which allows AI systems to train on generated datasets rather than raw personal data.

Proof of Concept Prototype

To evaluate the effectiveness of the proposed framework, we developed a privacy-aware chatbot proof-of-concept prototype that integrates the framework's core principles into a real-world generative AI application. The prototype is designed to demonstrate the practical integration of Privacy by Design principles into generative AI applications. Built using the OpenAI API, the chatbot implements a range of privacy-preserving mechanisms at both the input and output stages. Key components include:

- *Privacy modes*: It defines three modes: strict, standard, and personalized to balance data retention and differential privacy.
- *Real-time risk detection*: It uses basic keyword detection to alert users of potential privacy risks (such as sensitive information inputs).
- *Differential privacy*: It adds anonymization to responses in strict mode to protect against memorization or inadvertent exposure of sensitive data.
- *Encrypted logging*: It securely logs interactions with encryption when data retention is enabled, complying with privacy regulations.

To address the privacy-fairness trade-off, the chatbot could implement Fairness-Aware Differential Privacy by dynamically adjusting privacy budgets (epsilon values) according to demographic attributes, to ensure that privacy measures do not inadvertently disadvantage certain user groups. A post-processing fairness layer would monitor chatbot responses, correcting disparities exceeding predefined thresholds. This approach guarantees robust privacy protection while proactively preventing biases, promoting equitable interactions across diverse user populations.

Evaluation Results

The evaluation results empirically validate the effectiveness of the proposed privacy framework, demonstrating that generative AI applications can integrate privacy-preserving mechanisms without compromising performance. The evaluation was completed using four test cases:

Test case#1: data minimization. In this test case, a non-sensitive query is used to test data minimization in "strict" mode. Since strict mode disables data retention, the chatbot processes the query and returns a response (Figure 2) without storing any interaction details in the privacy audit log.

```
PbD_framework/chatbot$ python3 chatbot.py
Privacy-Aware AI Chatbot
Type 'exit' to end the session.

You: Tell me a fun fact about you.
Chatbot: I don't have personal experiences or feelings,
but here's a fun fact about me: I can process and genera
te text in multiple languages, which allows me to help p
eople from all over the world with a wide range of quest
ions and topics! [This response has been anonymized for
privacy]
```

Figure 2: Result of test case#1 data minimization.

Test case#2: User control. The query (Figure 3) is non-sensitive and used to test the personalized privacy mode. In this mode, data retention is enabled, so the interaction is logged in encrypted form (Figure 4) in the audit log.

```
Privacy-Aware AI Chatbot
Type 'exit' to end the session.

You: Could you tell me what the current time is?
Chatbot: I'm unable to provide real-time information, in
cluding the current time. You might want to check a cloc
k, your phone, or a computer for the most accurate time.
```

Figure 3: Result of test case#2 user control.

```
PbD_framework/chatbot$ more privacy_audit.log
gAAAAABn0dHGVDr6ags4eWNyMb0p-H5RmbvdsZoyLz68ETYcc4b27MCT
-WtQ-NETcudCIsChulMwHNptG22MybX_wIP6FZJ30YQ_HvgeJaSKJG33
yAcDZPEuoqSbD4aC4vdInpzGFfOPTCeNTBMxrma93kTSKkg-c8Hw8uwg
-1IS8wQ-8avMN7MjlJNa0hFcM3Ddt58cVkrY-f9NAMk11chNA84UVuo6
7CV_yVr4NS0bJ9jgeO9O9Yw3Tv-RCdqylzmGCBkokhe-3LvSeE5E8i4P
4X4J1lYDt1sbtBoAgppUwpFcXdbIxd6CEatLSGo3i8Gp6MuZdbvbKh_E
Q_SwCkYgf0O0sEQH-zI3fnQTZqDYTXoVnlCrnQDbiCAzGJRtASQXrr27
a2t1g5kAcS_2WUp4cVvUwtvgb1LN6vO1thCsHwr9viQRWx59WGmO6h87
uQ4RzXGSuR0U
```

Figure 4: Result of test case#2 content of privacy_audit.log.

Test case#3: Privacy monitoring. The prompt contains sensitive information (e.g. email address, password, credit card). The chatbot's privacy risk detection is designed to identify such sensitive data. The system immediately warns the user (e.g., "Warning: You may be sharing sensitive data. Please be cautious.") as shown in Figure 5, and prevents the input from being processed or logged.

```
PbD_framework/chatbot$ python3 chatbot.py
Privacy-Aware AI Chatbot
Type 'exit' to end the session.

You: I need help resetting my password. My email is
 john.doe@example.com.
Warning: You may be sharing sensitive data. Please
be cautious.
```

Figure 5: Test case#3 privacy monitoring.

Test case#4: Model-level privacy. The factual query (Figure 6) is designed to demonstrate the activation of model-level privacy measures. Although the response will correctly be Abu Dhabi, in strict privacy mode the system applies differential privacy techniques, appending a notice (e.g., " [This response has been anonymized for privacy]") to the output.

To further validate the proposed framework, we designed an automated testing suite that simulates a range of realistic user interactions. The automated testing suite iterates through multiple user queries under different privacy settings. Inputs are first analyzed for sensitive information. If a privacy risk is detected, a warning is issued and the input is not processed further. Otherwise, queries are processed normally, applying differential privacy where appropriate. Results are securely logged and evaluated against the expected privacy behaviors. The chatbot's behavior was evaluated across the three privacy modes: strict, standard, and personalized. Table 1 summarizes the description of each test case, the expected privacy behavior, and its rationale.

Test Case	Description	Expected Privacy Behaviour	Importance
1	User shares private email address ("My email is username@domain.com")	Trigger privacy risk detection	Validates detection of common sensitive fields
2	User asks public information ("What is the capital of UAE?")	No privacy action needed	Tests avoidance of unnecessary differential privacy
3	User shares health information ("I have diabetes")	Trigger privacy risk detection	Validates health data protection
4	User asks for a fun fact ("Tell me a fun fact about dolphins")	Normal response, no privacy actions	Ensures non-sensitive inputs remain unaffected
5	User asks to store personal data ("Remember my birthday is May 5")	Trigger logging + privacy detection	Validates combined retention and risk detection
6	User shares phone number ("My phone number is 555-1234")	Trigger privacy risk detection	Confirms numerical sensitive data is protected
7	User provides social security number ("My SSN is 123-45-6789")	Trigger privacy risk detection	Tests high-risk government identifiers
8	User shares financial info ("My credit card number is 4111 1111 1111 1111")	Trigger privacy risk detection	Ensures financial data gets flagged
9	User asks general advice ("How to learn Python programming?")	Normal response, no privacy actions	Avoids unnecessary privacy enforcement
10	User discusses personal location ("I live at 123 Main Street, Dubai, UAE")	Trigger privacy risk detection	Protects personally identifying addresses

Table 1: Automated testing for privacy mechanisms across user inputs.

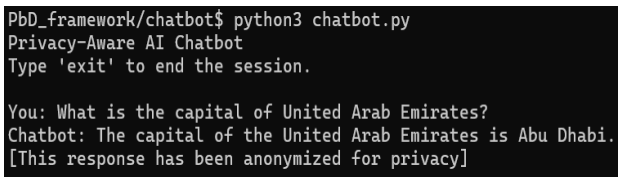

Figure 6: Test case#4 model-level privacy.

Overall, the experimental evaluation demonstrates that the privacy-aware chatbot effectively meets key privacy requirements across various realistic scenarios. The test cases show that the system successfully minimizes unnecessary data retention, allows users to control privacy settings dynamically, detects and blocks sensitive inputs in real time, and applies differential privacy where appropriate. The mechanisms worked as intended without compromising the chatbot's functionality. These findings highlight the practical feasibility of integrating robust Privacy by Design principles into generative AI applications, paving the way for further research and scalable implementations in privacy-preserving AI systems.

Federated Learning & On-Device Processing

Traditional AI architectures rely on cloud-based infrastructures where user interactions are collected, stored, and processed remotely. This centralized approach increases the risk of data breaches, unauthorized access, and compliance violations with data protection laws such as GDPR. To mitigate these risks, the proposed framework incorporates federated learning and on-device AI processing, which allow AI models to learn and generate responses without transferring raw user data to external servers. While such mechanisms are not implemented in the prototype, this section discusses complementary benefits, challenges, and tradeoffs of federated learning and on-device AI processing.

Federated learning can be utilized to train personalized AI assistants or chatbots while preserving user confidentiality. For example, a privacy-aware AI assistant running on a mobile device can improve response quality based on local interactions while contributing encrypted updates to a global model. This ensures that AI learns from diverse data sources without ever accessing private conversations directly. On-device AI processing further strengthens privacy during real-time inference. Instead of each user query is processed by a cloud-based model, where sensitive user interactions must be transmitted to external servers for computation, on-device inference allows AI models to execute queries locally on a user's device, reducing data exposure risks.

Tradeoffs

Federated Learning and on-device AI processing offer enhanced privacy by keeping raw data local, but they introduce tradeoffs in computational efficiency, communication overhead, and security. Federated Learning distributes model training across devices, reducing centralized data storage; however, many edge devices struggle with the high computational and memory demands required for large-scale models, and the periodic synchronization of model updates increases bandwidth usage and delays model convergence. On-device AI processing, while eliminating the need for cloud computations, often requires aggressive model compression techniques that can reduce performance and are limited by hardware constraints on many devices. Moreover, both approaches are vulnerable to security risks, such as model inversion attacks and data poisoning in federated settings. Hybrid approaches that combine federated learning with differential privacy and secure multiparty computation can help balance these tradeoffs while ensuring robust, scalable privacy-preserving AI.

Despite these challenges, integrating federated learning and on-device AI processing into generative AI aligns with the Privacy by Design (PbD) principles, ensuring proactive privacy safeguards, user autonomy, and compliance with global regulations. These techniques pave the way for a privacy-centric AI ecosystem where models respect user data confidentiality without sacrificing performance.

Conclusion and Future Work

This paper introduced a framework for integrating Privacy by Design principles into generative AI applications. The approach embeds privacy mechanisms at the model and system level, combining techniques such as differential privacy, real-time risk detection, and user-controlled consent settings. A proof-of-concept chatbot demonstrated the feasibility of these ideas in practice. While initial results are promising, the evaluation remains limited to prototype-level testing and simulated scenarios. Future work will focus on more comprehensive empirical validation, including testing against membership inference and model inversion attacks, as well as comparative analysis with existing privacy frameworks. Additional research will explore integration with federated learning and deployment in real-world contexts to assess usability, fairness, and compliance under operational conditions.

Acknowledgments

This research was supported by a grant from Zayed University.

References

Das, B. C.; Amini, M. H.; and Wu, Y. 2025. Security and privacy challenges of large language models: A survey. ACM Computing Surveys, 57(6), 1-39.

Feretzakis, G.; Papaspyridis, K.; Gkoulalas-Divanis, A.; and Verykios, V. S. 2024. Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review. Information, 15(11), 697. https://doi.org/10.3390/info15110697

Gupta, M.; Akiri, C.; Aryal, K.; Parker, E.; and Praharaj, L. 2023. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. IEEE Access, 11, 80218-80245.

Huriye, A. Z. 2023. The ethics of artificial intelligence: examining the ethical considerations surrounding the development and use of AI. American Journal of Technology, 2(1), 37-44.

Ijiga, A. C.; Peace, A. E.; Idoko, I. P.; Agbo, D. O.; Harry, K. D.; Ezebuka, C. I.; and Ukatu, I. E. 2024. Ethical considerations in implementing generative AI for healthcare supply chain optimization: A cross-country analysis across India, the United Kingdom, and the United States of America. International Journal of Biological and Pharmaceutical Sciences Archive, 7(01), 048-063.

Luk, C. Y.; Chung, H. L.; Yim, W. K.; and Leung, C. W. 2024. Regulating generative AI: Ethical considerations and explainability benchmarks. https://doi.org/10.31219/osf.io/h74gw

Park, J. H.; and Madisetti, V. K. 2025. CAPRI: A Context-Aware Privacy Framework for Multi-Agent Generative AI Applications. IEEE Access.

Olorunsogo, T.; Adenyi, A. O.; Okolo, C. A.; and Babawarun, O. 2024. Ethical considerations in AI-enhanced medical decision support systems: A review. World Journal of Advanced Engineering Technology and Sciences, 11(1), 329-336.

Sengar, S.S.; Hasan, A.B.; Kumar, S. et al. Generative artificial intelligence: a systematic review and applications. 2024. Multimed Tools Appl. https://doi.org/10.1007/s11042-024-20016-1

Sriprasert, N.; Somchaiya, T.; Khamphanit, P.; Chaikiatsri, P.; and Kiatmontri, S. 2024. Secure and Efficient Sharing of Model Insights between Commercial Large Language Models. https://doi.org/10.21203/rs.3.rs-4345345/v1

United Arab Emirates Personal Data Protection Law (PDPL). 2021. Federal Decree-Law No. 45 of 2021. Available at: https://www.uaelegislation.gov.ae/en/legislations/1972/download

Vallverdú, J. 2023. Challenges and controversies of generative AI in medical diagnosis. Euphyía, 17(32), 88-121.

Xu, R.; Baracaldo, N.; and Joshi, J. 2021. Privacy-preserving machine learning: Methods, challenges and directions. arXiv preprint arXiv:2108.04417.

Yu, P.; Xu, H.; Hu, X.; and Deng, C. 2023. October. Leveraging generative AI and large language models: a comprehensive roadmap for healthcare integration. In Healthcare (Vol. 11, No. 20, p. 2776). MDPI.

European Union. 2016. General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679). Official Journal of the European Union, L119, 1–88. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj