BAFFLE DATA PROTECTION FOR AI

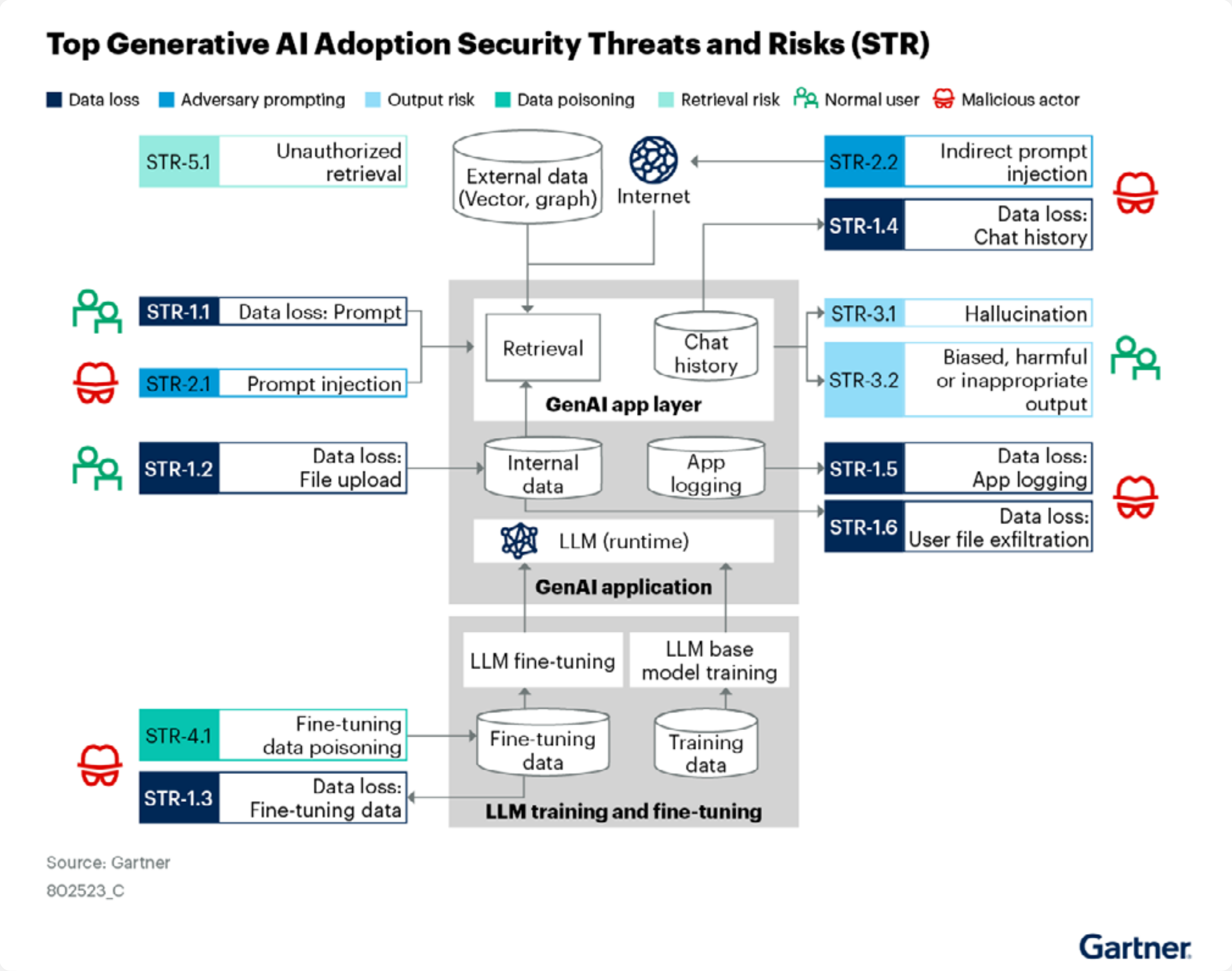
The Easiest and Most Effective Way to Secure Sensitive Data used in Generative AI (GenAI) Applications

Companies around the world are in the process of using GenAI to increase revenue potential and decrease costs for their business. To do so, they need to move away from GenAI applications that use public or non-confidential data and start incorporating private and sensitive data into the GenAI pipeline. To help enterprises enable this new generation of AI application, Baffle provides a solution that ensures the data used by GenAI applications is secured and in compliance with privacy regulations.

Data Security Challenges in GenAI Applications

Deploying and using GenAI applications creates new security risks. As evident from a recent Gartner report on generative AI adoption security threats, the prevalent risk involves data. This is hardly surprising given that:

- AI applications store and process vast amount of data to be effective
- A full GenAI application pipeline introduces a myriad of new repositories where data is cached or stored.



Top Generative AI Adoption Security Threats and Risks (STR)

Data loss
Adversary prompting
Output risk
Data poisoning
Retrieval risk
Normal user
Malicious actor
STR-5.1 Unauthorized retrieval
External data (Vector, graph)
Internet
STR-2.2 Indirect prompt injection
STR-1.4 Data loss: Chat history
STR-1.1 Data loss: Prompt
STR-2.1 Prompt injection
Retrieval
Chat history
GenAI app layer
STR-3.1 Hallucination
STR-3.2 Biased, harmful or inappropriate output
STR-1.2 Data loss: File upload
Internal data
App logging
STR-1.5 Data loss: App logging
STR-1.6 Data loss: User file exfiltration
LLM (runtime)
GenAI application
LLM fine-tuning
LLM base model training
STR-4.1 Fine-tuning data poisoning
Fine-tuning data
Training data
STR-1.3 Data loss: Fine-tuning data
LLM training and fine-tuning
Gartner

Source: Gartner
802523_C

The challenges of securing sensitive data are compounded by existing data security and privacy compliance requirements that require entities to be responsible for the regulated data that they control and that appropriate security mechanisms are put in place to ensure that the regulated data is only made available on an “as-needed” basis to those who require access.

Baffle Addresses All GenAI Data Security Challenges

Cybersecurity vendors have taken a variety of approaches to secure GenAI applications, but none have been able to ensure that sensitive and regulated data can be protected in a manner that meets existing data security and compliance requirements. Approaches such as malicious prompt detection and filtering are readily bypassed with adversarial prompting, and existing data loss prevention tools result in false positives and negatives that allow sensitive data to slip through while adding additional monitoring overhead.

Baffle Data Protection for AI provides field-level encryption that protects sensitive data values used by GenAI applications for context information and in training models. Baffle only allows those authorized by the access control policies the ability to see these values. With Baffle, organizations adopting GenAI can demonstrate to their compliance auditors the use of verifiable security controls to protect regulated data used for GenAI.

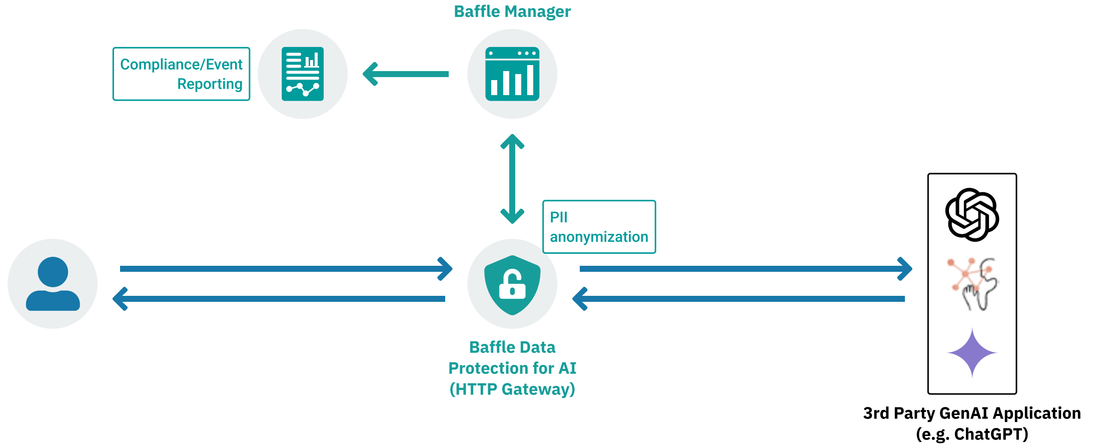
Key Benefits

- **Secure:** Regulated personally identifiable information (PII) is encrypted and remains protected and controlled throughout the GenAI application infrastructure
- **Easy:** No application changes required to protect data cryptographically everywhere it is used
- **Verifiable:** Fully auditable role-based access control policies deterministically ensures that only authorized users have access sensitive data
- **Flexible:** Baffle Data Protection for AI is delivered as fully containerized software that can readily scale horizontally to meet all types of workloads

Baffle Provides Data Protection for All Stages of GenAI Adoption

Enterprises are taking a staged and heterogenous approach to deploying and using GenAI applications. Baffle's Data Protection for AI provides comprehensive data security capabilities that addresses the risk for the GenAI application usage scenarios at each stage of the journey.

Stage 1 - 3rd party GenAI SaaS

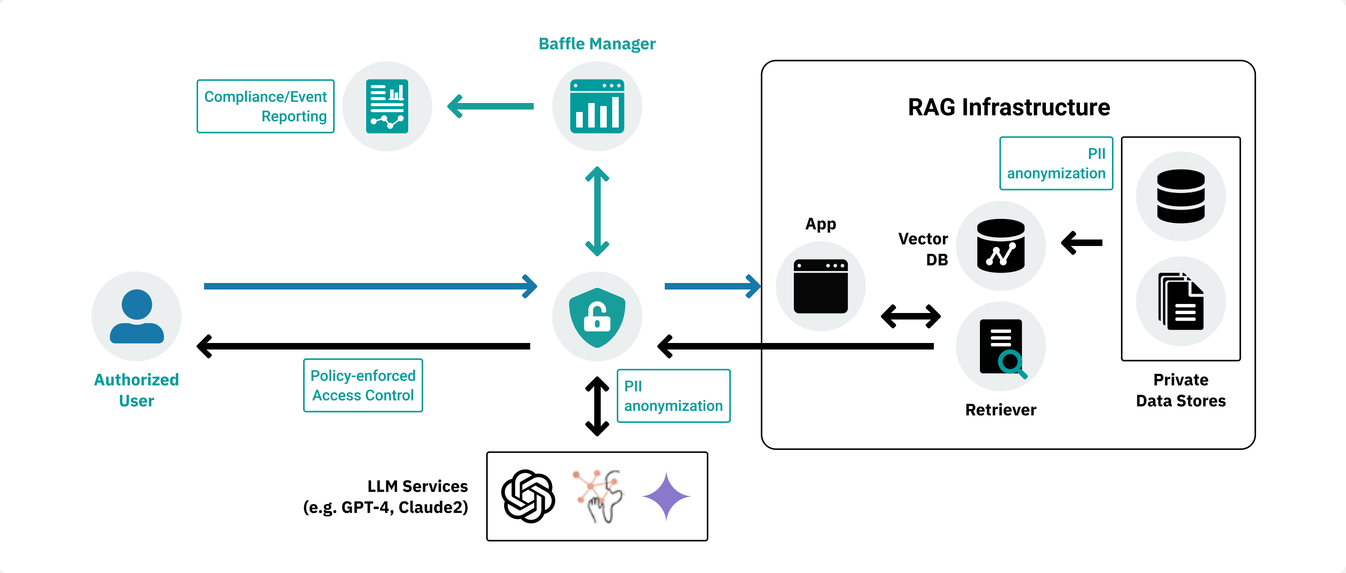

Baffle Manager
Compliance/Event Reporting
PII anonymization
Baffle Data Protection for AI (HTTP Gateway)
3rd Party GenAI Application (e.g. ChatGPT)

For most companies, their first GenAI application is often a fully encapsulated SaaS application. For example, many organizations currently use ChatGPT for text generation, search, and summarization assistance without deploying any GenAI infrastructure or implementing GenAI application code. For these types of deployments, the primary data security risk comes from internal users sending sensitive data values to these 3rd party applications.

Baffle provides an HTTP security gateway that addresses the data security risks associated with using 3rd party GenAI applications and services. This security gateway identifies PII values on the fly using a named entity recognition and classification (NERC) process based on the latest available AI models. It then encrypts these values before they are sent to the SaaS provider.

By encrypting only sensitive and regulated PII, the GenAI service can make use of all non-sensitive data for inferencing and processing while remaining ignorant of the sensitive values that they should not possess.

Stage 2 – AI Application Using Retrieval Augmented Generation



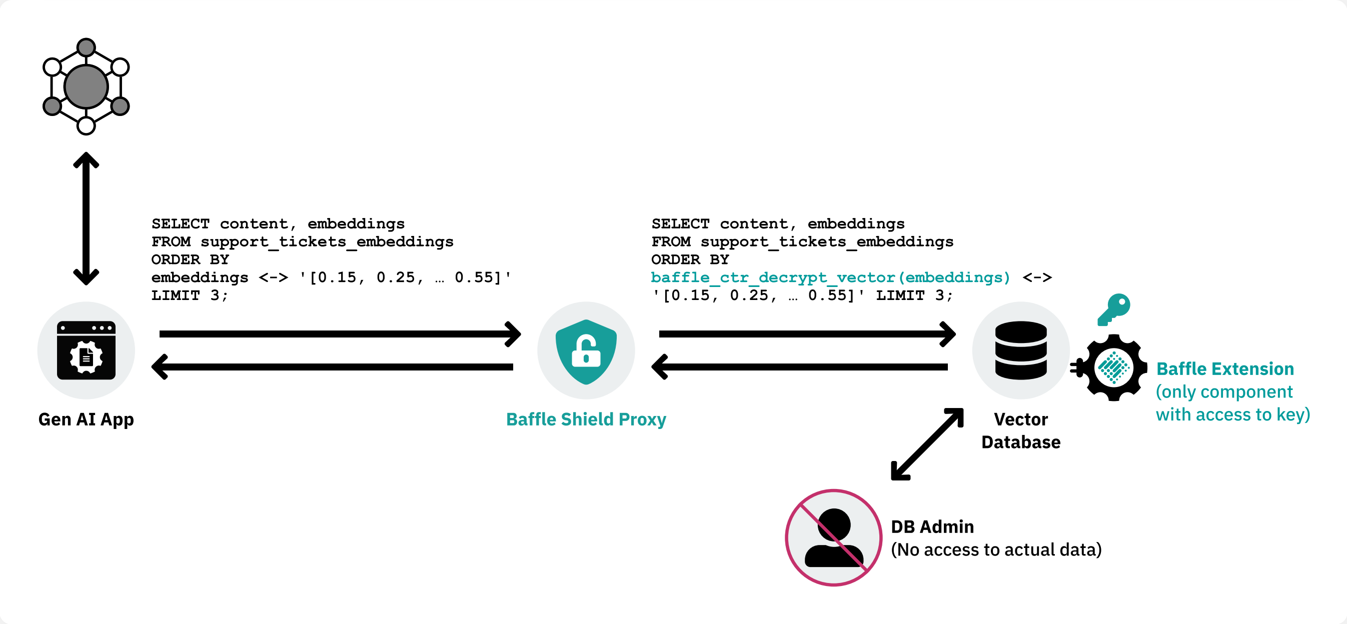
For organizations that need more custom capabilities from their GenAI application or need to tap into a large set of private and sensitive data, developing a GenAI application using retrieval augmented generation (RAG) is a common approach. RAG allows users to search, summarize, and generate new content using off the shelf large language models in conjunction with an existing body of contextual data. In most cases, the data used for context need little or no preparation or processing for use by a RAG application, and the responses from a RAG application are much less prone to hallucination due to the use of contextual data to limit the scope of the response.

Baffle can encrypt data and documents in popular structured and unstructured data stores at a field level without changes to the GenAI application or developing new tools. This ensures that sensitive or regulated PII in context documents used by GenAI applications using RAG is fully protected in all parts of a RAG Gen AI application including:

- In the vector database
- In API calls for GenAI inferencing
- In responses to end users

Secured Use of Vector Databases

To secure vector databases for use by GenAI applications, Baffle provides a reverse proxy that transparently encrypts sensitive data values stored in the database including vector embeddings. While encrypted, GenAI applications can perform operations such as similarity search on the encrypted embeddings without decrypting them. Baffle accomplishes this by automatically injecting the function calls to use the Baffle database extensions to operate on encrypted values when it detects applications calls to perform database operations on encrypted data.



```
SELECT content, embeddings
FROM support_tickets_embeddings
ORDER BY
embeddings <-> '[0.15, 0.25, … 0.55]'
LIMIT 3;
```

```
SELECT content, embeddings
FROM support_tickets_embeddings
ORDER BY
baffle_ctr_decrypt_vector(embeddings) <->
'[0.15, 0.25, … 0.55]' LIMIT 3;
```

Gen AI App
Baffle Shield Proxy
Vector Database
Baffle Extension (only component with access to key)
DB Admin (No access to actual data)

Currently, the ability to encrypt vector embeddings and perform operations on encrypted embeddings is only available when using PostgreSQL with pgvector as the vector database. Baffle has plans to extend this capability to other vector databases in the future.

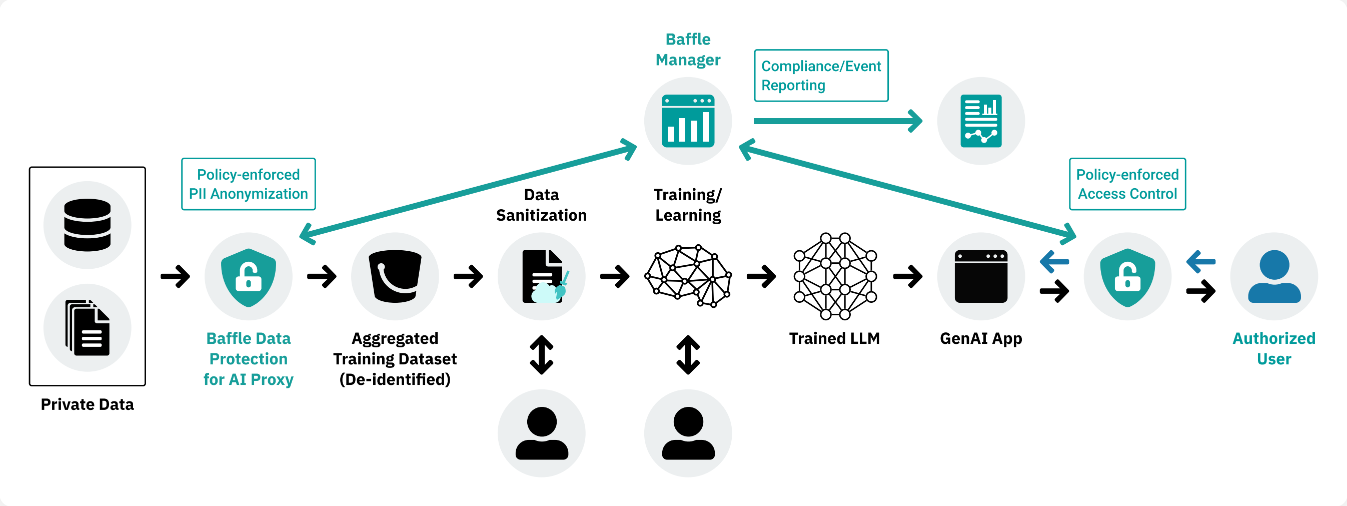
Secured Inferencing

To secure sensitive data in API calls used for inferencing, Baffle offers the ability to encrypt the data values at a field-level documents   used as context in the inferencing calls. Baffle also provides a security gateway for API calls to 3rd party inference servers that will automatically detect and encrypt known PII values to prevent them from being exposed to 3rd parties. The detection uses Baffle's unique multi-model AI-based named entity recognition technology that uses multiple AI models for detecting sensitive values.

Secure Responses

To secure responses from GenAI applications to end user requests and prevent them from accessing information they otherwise would not have access to, organizations can use Baffle Data Protection for AI solution that includes an HTTP proxy that will decrypt or mask sensitive data values based on predefined access control policies. Furthermore, these access control policies can leverage roles and attributes defined in any Open ID Connect (OIDC) identity provider through claims in the Java web token (JWT). The result is a set of verifiable controls that can limit user access to all, part or none of the sensitive data values for security and compliance.

Stage 3 - Training/Fine-tuning New Large Language Models (LLM)


Baffle Manager
Compliance/Event Reporting
Policy-enforced PII Anonymization
Data Sanitization
Training/ Learning
Policy-enforced Access Control
Private Data
Baffle Data Protection for AI Proxy
Aggregated Training Dataset (De-identified)
Trained LLM
GenAI App
Authorized User

Some companies require more customized capabilities from their GenAI applications. To meet their requirements, they will train a new AI model or fine-tune an existing AI model with their own private data. This data will often include sensitive or regulated PII that will be exposed to multiple parties during the training process and in responses to users of the model.

While new software tools have simplified the training process at a technical level, training high quality AI models still requires a large amount of good, clean data and some supervision in the training process. These requirements introduce new data security risks in addition to those from using a GenAI application. These include:

- Administrators and users of the data store where all of the training data is aggregated from multiple sources
- Administrators and users of the systems involved in sanitizing the training data
- Administrators and users of the systems involved in the supervised training of the model

Baffle Data Protection for AI addresses the risks of sensitive data exposure when training or fine-tuning new models by enabling the encryption of sensitive and regulated PII values at a field-level with minimal effort. This ensures that the data is never accessible by unauthorized users as it moves through the GenAI training pipeline.

Learn More

Get personalized insights and recommendations from Baffle's GenAI data security experts for your current or upcoming GenAI projects. Schedule a consultation here.

info@baffle.io
https://baffle.io
3979 Freedom Circle, Suite 970
Santa Clara, CA 95054