

Large Language Models (LLMs) & The Implications for Privacy & Security Teams





What Inspired the Rise of LLMs and Why Now?

We have witnessed the remarkable rise of generative AI, powered by vast amounts of pre-trained data within large language models (LLMs). Chat-based interfaces have democratized AI for the general public, which has led us to where we are today.

The underlying LLM technology enables anyone to use prompts using natural language to elicit queries for a specific or intended response. These responses can range from generating text, code, performing tasks, and more. The “natural language understanding” aspect is similar to humans, which previously has not been possible.

We have seen chatbots serve as amazing and effective stock market recommenders, customer support agents, legal case analysts, and even healthcare assistants for complex documents with sophisticated personalities and conversations while retaining context to complete transactions. The current set of conversational AI is already helping businesses scale with routine tasks like generating invoices, automating returns, researching complex topics, building initial content, and more, saving hours and even days of human work. We are seeing the initial impact of how far automation can be leveraged and the intensity of complex tasks that can be handled with AI. Given these cost savings and efficiency, we will see a widespread use of AI chatbots, agents, and task handlers continue to increase.

With these capabilities, customer data can be easily misused as private conversations. Activity and logs can now be combined with user cookies to allow privacy and security breaches within a simple prompt. As we do not yet have data governance controls for LLMs, a malicious actor can request and receive sensitive data related to a business or person with relative ease.

How LLMs Evolve with Reinforcement Learning with Human Feedback

LLMs need to be trained on large amounts of data including text, audio, and video. Human input is needed to provide feedback for fine tuning the responses from LLMs or provide a baseline for correct and incorrect answers. Current LLMs can be trained over hundreds of billions of text documents covering a wide range of situational information, creating a knowledge base that can answer a wide range of questions. This input data set varies and consists of source code, database schema, sales metrics, customer demographics, legal contracts, IT support tickets, patient chats, slack messages, etc.

Today, 80-85% of data is unstructured, and many businesses lack tools to fully utilize this data type. With LLMs, we have an opportunity to effectively make sense of all of this rich unstructured information and can streamline efforts to use it for meaningful purposes.

Due to auto-learning capabilities, LLMs can be constantly fed data. The amazing memorization capabilities of LLMs provide us with querying, recommendation, and problem-solving capabilities, which have not been possible before. One of the key capabilities offered by LLMs is that of feedback-based learning where responses can be augmented and improved over time.

How Businesses Interact with LLMs Today

Business interacts with LLMs in three ways today:




1. LLM-hosted platforms directly using an LLM created and maintained by businesses with AI expertise, such as OpenAI.
2. Embedded apps via chat/conversational bots within a currently used platform like Google Docs or Office365

3. Self hosting model - Either train an LLM from scratch or utilize an Open Source LLM like Alpaca, fine tune the weights and maintain a self-hosted version.

As of now, few companies have the cloud infrastructure and AI expertise to create these models from scratch and manage/maintain them for others to consume. Several cloud providers including Amazon Web Services, Google Cloud Platform, and Microsoft Azure are in the process of offering host LLM based services. Smaller models that can be easily trained and run on a restricted compute environment including mobile devices are also being actively researched.





Today, most businesses use Embedded Application Workflows to interact with LLM prompts within their day-to-day. Tools include word processors, source code checkers, SQL query generators, email responders, customer support, meeting summarizers, trip planners, legal document analyzers, and more. Through these approaches, workflows/bots are able to ingest custom datasets and provide responses against them.

Privacy/Security Concerns Introduced via LLMs

There are several data security and privacy concerns introduced with LLM usage in a business context. Our intention with this paper is to provide an overview with guidance towards mitigation or remediation.

The key areas are:

Biased Outputs

Businesses need to be vigilant about using LLMs for activities prone to biases e.g., analyzing resumes for employment fit, automating customer service needs for low-income vs high income groups, or forecasting healthcare issues based on gender/age/race. A major issue in AI data training today is due to unbalanced data where one category of data is overwhelmingly dominating other categories leading to bias or incorrect correlations. A typical example would be any dataset with race, age, or gender distributions. Any kind of unbalanced data in these areas could lead to unexpected, unfair outcomes. And if the LLMs are trained by third parties the degree of bias due to these factors is unknown to the LLM consumer.

Explainability & Observability Challenges

For the current set of LLMs hosted publicly, there are only a few prompts available to tie in output results to known input. LLMs can “hallucinate” to create imaginary sources, making observability a challenge. For custom LLMs, businesses can inject observability during training to create associations during the training phase of an LLM. And then it would be possible to correlate the answers to that of the underlying sources to validate the output. Businesses need to set up bias measurement and monitoring to ensure that the output of LLMs does not lead to harm or discrimination in these scenarios. Imagine harm caused by a LLM-based medical notes summarizer producing different health recommendations for males versus females.



Privacy Rights & Auto-Inferences

As LLMs ingest data, they can create inferences with any personal information categories being provided as customer service records, behavior monitoring, or products considered. Businesses need to ensure that they have appropriate consent as a processor or sub-processor to derive these inferences. It is incredibly hard and expensive for businesses to keep track of privacy data rights and restrict usage in the current setup.

Unclear Data Stewardships

Currently there are no easy, efficient ways for LLMs to unlearn information. The way businesses are using sensitive data as processors or sub-processors makes data stewardship complex to manage. This increases the legal obligations significantly for businesses. For security teams, data inventory, classification, and automation is crucial to design adequate safeguards for AI systems input and output responses. Input data into LLMs for training or prompts need to be filtered to ensure that information used is identified within the scope, for the purpose of use.


Data Discovery for Large Language Models (LLMs)

Around 2008, storage prices reached a **downward** inflection point and have been drastically reducing since then. Technological advances, such as the rise of SaaS, along with declining storage prices have made it quite easy to collect unstructured data from multiple sources including customers, data brokers, and sub-processors.

And businesses are storing this data at far greater levels than before to derive meaningful financial results.

Data being harnessed today can be categorized in these four buckets today:

- **Structured:** Standard databases (e.g. Oracles or MySQL) where data types and classifications are defined with strict definitions and are relatively easily discoverable via SQL based tooling
- **Semi-Structured:** Excel files, Graph Databases, CRMs like Salesforce, Servicenow, Infrastructure Configuration with shared secrets, etc.
- **Unstructured:** Documents, emails, files, Source code, Contracts, DPAs, etc.
- **Multimedia:** Audio, Video, Moving Gifs, 3D Autocad Drawings, AR-VR Data, etc.




It's critical for businesses to make it convenient for users to buy and search for products through multiple touchpoints ranging from email, chat, phone, or text messages. This mixed-mode use of data collection makes governance a tedious process given that there is an exponential rise in unstructured data usage. About 80% of data collected today is unstructured and is the fastest growing category containing sensitive data.

Since LLMs can ingest multi-modal datasets structured and unstructured, a NextGen AI is needed for multi-dimensional data discovery and correlate data attributes across multiple sources.

NextGen AI for contextual data discovery can provide an accurate picture of what sensitive data including PII, IP or Confidential information is available in email, documents, data stores, applications, and more. Our self-learning, adaptive AI platform creates a lineage of all attributes across multiple sources and correlates it with Individuals/Businesses/Vendors/Process/Risks. The self-learning AI approach eliminates manual creation of rules or policies to classify or identify any data categories. Once the data mapping is complete and inventoried, Users' get a comprehensive view of their data crawl spread across sources.

Once the data inventory is completed, Privacy & Security teams will get a 360-degree view to study the scope of security/privacy & legal risks to decide which data sources, documents, emails, databases can be shared/used for LLM based services.



In cases, which require LLMs be exposed to sensitive personal data categories, Users can strategize to anonymize the sensitive data with generic tokens with anonymized IDs as “John” with “Person_12”, “SSN_Number” with “SSN_DATA_10” to protect individuals and prevent bias. Masking helps create a “balanced” and correlated dataset for improved accuracy and sensible results. As long as the data lineage is intact it would still be useful for a wide variety of purposes.

Other guardrails for preventing data usage with LLMs would need custom tagging & access control. We will discuss these in our next series of blogs.

Why Self-Learning AI for data discovery: 80-85% of data is unstructured today. It is impossible, tedious and error-prone to create rules/policies around sensitive data identification & usage. Businesses are JIT today with multiple software releases every week. SaaS applications can start collecting information rapidly from multiple sources making it hard for legal/security/privacy teams to effectively implement data governance. We believe this approach helps to focus on their core priorities rather than writing static rules or policies which provide a limited view of sensitive data missing associative attributes.

Why is Data Discovery Critical Now?

Large Language Model (LLM) prompts now provide an easy mechanism to combine and query unstructured, semi-structured, and unstructured data capabilities to organizations. As LLMs have memorization and contextual linkage capabilities for multiple data categories, sensitive data can easily be queried and misused for a variety of purposes. Privacy, security, and legal teams need to have access to a data discovery tool that can search and correlate sensitive data across multiple categories. Once the data is found, data governance can be implemented.


The Contextual Awareness and Data Inference Capabilities of LLMs Allow:

- User Attributes can be derived easily from partial information - For example, a phone number can be extracted by prompting information - Tell me how many users live in San Francisco? Which of these users are male and have age greater than 30?
- Bias Introductions - If underlying loan qualifier tickets data favor particular races, age genders, or zip codes, LLM analytics related queries for vetting new customers will be biased.
- Information Misuse and Correction - LLMs do not offer straightforward mechanisms to update or correct underlying data making it extremely hard to prevent data misuse; e.g If SSN information is accidentally ingested to an LLM we do not have mechanisms to prevent the sharing and misuse of this information. Further malicious users can combine personally identifiable information such as a social security number, birthdate, and name and address information to do broader financial harm, such as applying for new credit card accounts or claim social security benefits
- Nonexistent Data Retention/Erasure Capabilities - Current LLMs store word vectors that have thousands of associated/related keywords and attributes linked with any sensitive data. It is nearly impossible today to query and delete this data for a specific user or group of users.

- Lineage for Sensitive Data Results - No good tools exist today to provide details about how the data has been derived or calculated. The amount of data vectors and multiple dimensions make it computationally expensive to offer data lineage for sensitive data to give some sense of scale. The word “Cat” in Wikipedia has a vector representation of 300 in commonly available models. This implies that there are 300 associations with other words. To update/delete or take a look at the lineage of this word in Wikipedia, this is a lot of information to compute and study. Further, these 300 associations might have hundreds of other associations of their own, making it exponentially hard for all purposes.

Current Data Discovery Mechanisms for above Data Categories:

- Unstructured Data: Cybersecurity tools are mostly pattern-based or use minimal AI to identify/classify sensitive data in unstructured sources such as logs, emails or on the network.
- Structured/Semi-Structured Data: Data governance platforms offer visibility and insights for sensitive data in structured sources as databases or partially semi-structured data for data warehouses such as Snowflake or MongoDB. Popular data governance tools offer policy-based masking capabilities or access restrictions for sensitive data. Partial discovery and anonymization of a few data attributes will not work for LLMs as users can still derive relationships, e.g., prompting for phone number based on secondary information such as a zip code, location, or past employer.



Essentially, businesses need a comprehensive view for each individual's information to prevent any data misuse via LLMs throughout all sources within the company. Gone are the days where data can be encrypted and masked in a database, presuming that user's identity information cannot be derived by secondary attributes like address, zip code, income range, or even the model and make of car.