

The Boring Parts of AI

A photograph of a brown-throated three-toed sloth sitting at a wooden desk. The sloth is leaning forward, its front paws resting on a laptop keyboard. To the left of the sloth is a white computer monitor displaying a user interface with various icons and data. In the foreground, there's a white mug on the left and a small potted plant on the right. The background shows a window with a view of a city skyline at dusk or night. The overall mood is relaxed and humorous, suggesting that the "boring parts" of AI involve tasks like data governance, security, and privacy.

LLMs Data Governance,
Security & Privacy

Dan Fernandez

- Data Nerd
- Product Manager in Cybersecurity
- Instructional Associate @ GaTech



[@danielfernandez](https://twitter.com/danielfernandez)



[/in/dafdz](https://www.linkedin.com/in/dafdz)



danielfernandez@infosec.exchange



danielfernandez.medium.com



hey@dafnz.com

Disclaimer

- Opinions are my own and not my employers.
- Suggestions and considerations around data privacy are not legal advice.



Presentation Resources



- The slides will be made available in Github immediately after the presentation.
- Slides contain references to content with additional information.



Resource: [Will look like this](#)

Agenda

1. Intro to LLMs
2. Data Governance
3. Data Security & Privacy
4. ML Governance & Security
5. Securing ML Apps



Intro To LLMs



What are Large Language Models?

- A type of artificial intelligence (AI) that can generate text content by predicting word sequences.
- Trained on massive datasets of text and code.
- Becoming increasingly powerful, and are being used in a wide range of applications.



💡 Resource: [Why ChatGPT Works - Stephen Wolfram Writings](#)

Illustration by: MidJourney

Brief History

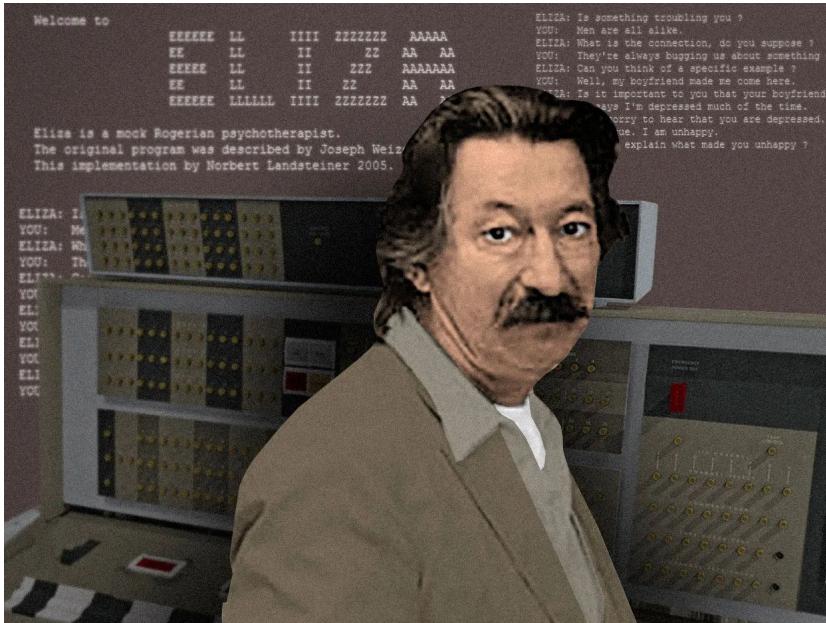
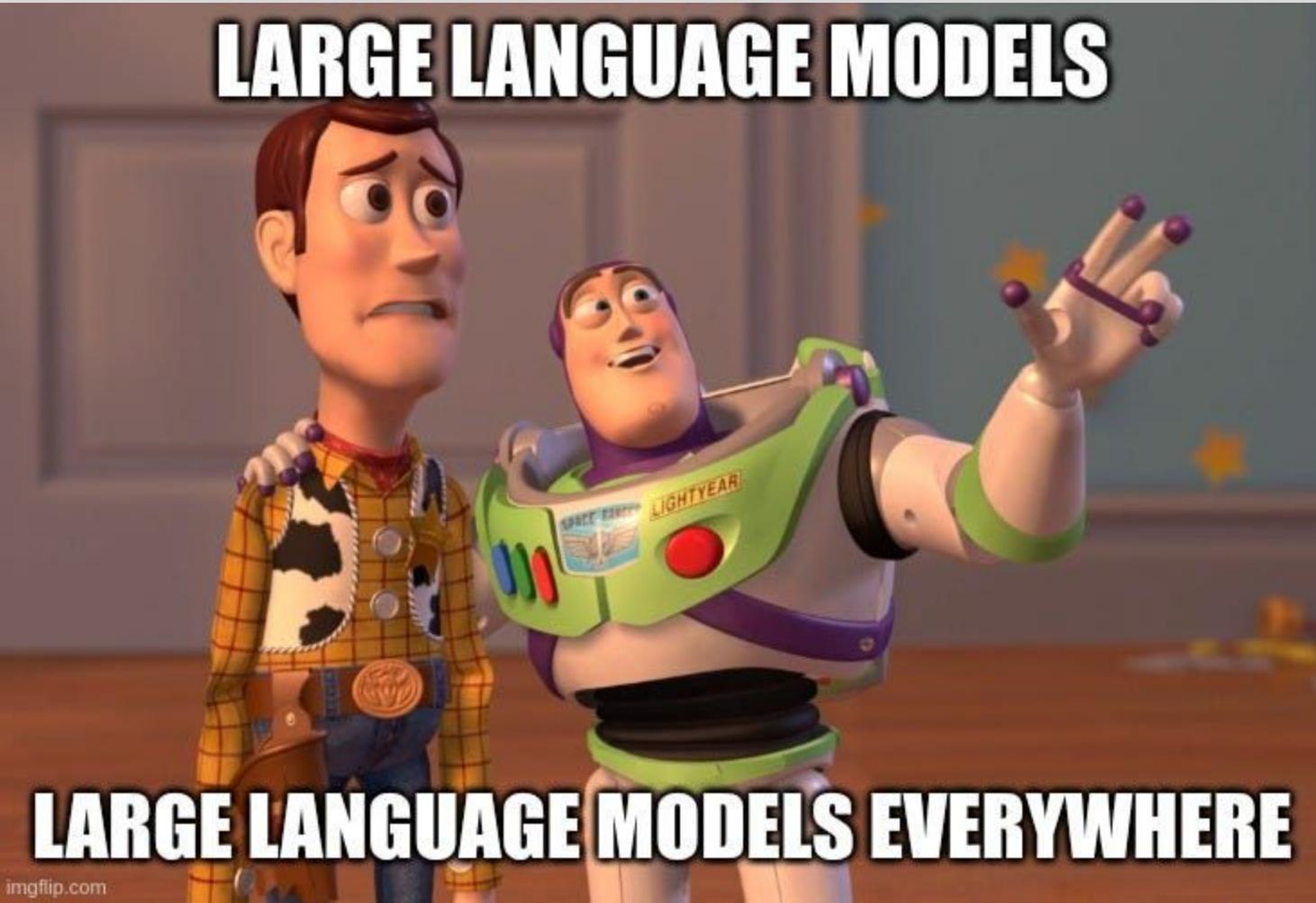


Photo Credit IEEE Spectrum: Joseph Weizenbaum

- 1950s - 1960s: First rule-based LLMs developed.
- 1980s: Statistical LLMs introduced.
- 2010s: Deep learning LLMs introduced.
- 2020s: LLMs trained on massive datasets and capable of performing a wide range of tasks.

LARGE LANGUAGE MODELS



Examples of Large Language Models

- GPT
- BERT
- Llama



OpenAI

Meta

Google



Resource: [A comprehensive view of Large Language Models](#)

LLM Use Cases

- **Text generation:** News articles, blog posts, and product descriptions.
- **Question answering:** Natural Language Answers
- **Summarization:** Index and summarize data
- **Code generation:** Software development
- **Translation:** Large Scale Language Translations



💡 Resource: [What are large language models used for? Nvidia](#)

Business Examples and Benefits



- **Security:** Query, analyze and understand security data
- **Customer service:** Chatbots that can answer customer questions and provide support.
- **Marketing:** Personalized marketing campaigns and generate content that resonates with customers.
- **Finance:** Analyze financial data.
- **Healthcare:** Personalize patient care.
- **Education:** Personalized learning experiences and help students learn more effectively.
- **Law:** Analyze legal documents and help lawyers with their research.



Resource: [10 LLM Project Ideas - Towards Data Science](#)

Data Governance



The need for Training LLMs

Why do it?

- ***Maximize Value*** of Company's Data
- ***Improve Accuracy*** and Performance for Custom Use Cases
- ***Enhance Security*** Features to the model

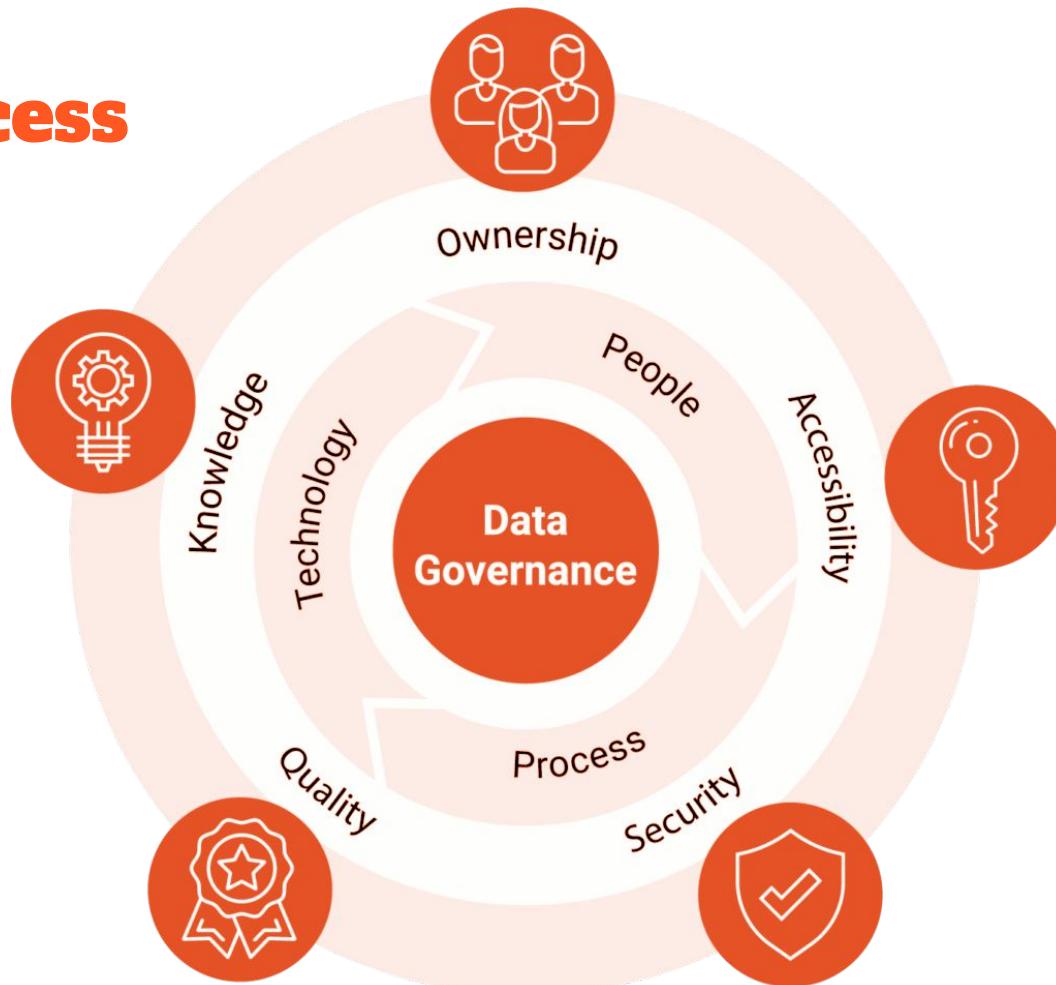
Training Options

- Custom LLM
- Fine Tuned LLM
- Prompt General Purpose LLM

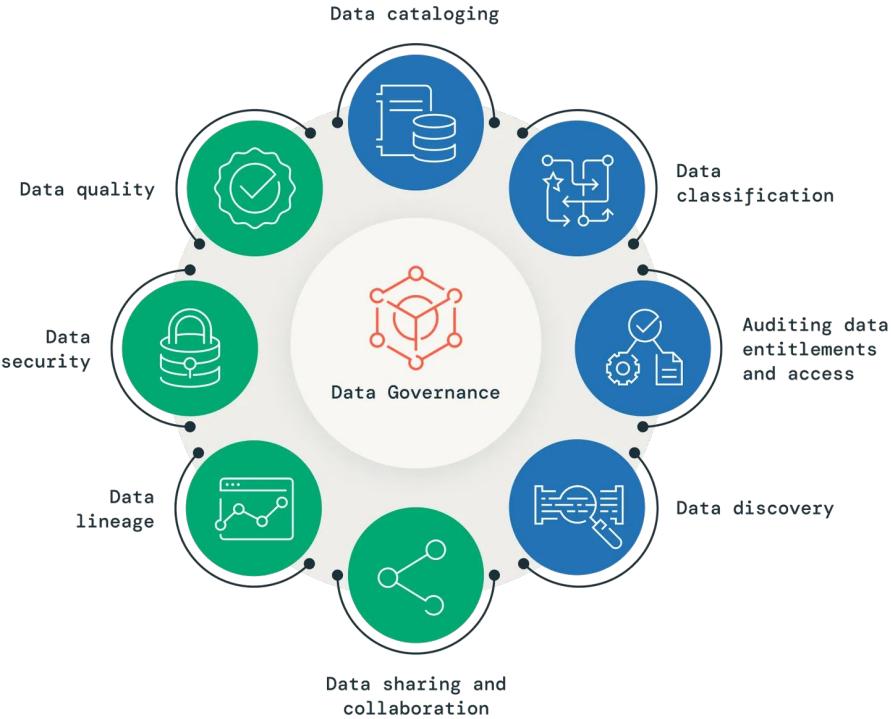


Illustration by: MidJourney

The Process



Data Governance Components



💡 Resource: [Data Governance Comprehensive Guide - DataBricks](#)

Data Governance Challenges



More Data = More Problems

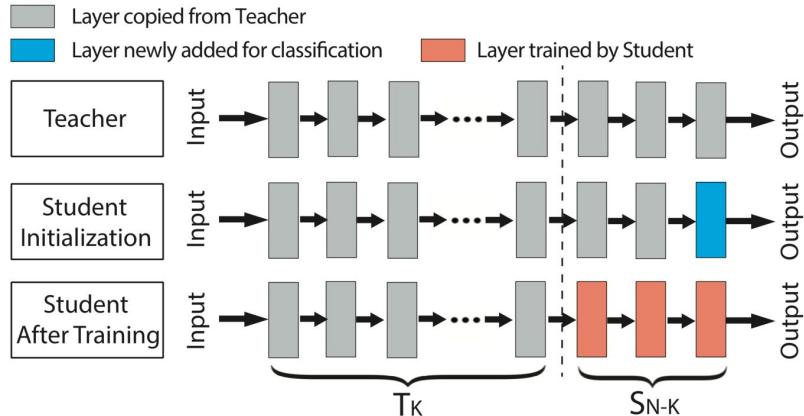
- Complex & Siloed Data
- Ethical Concerns
- Regulatory Compliance

Data Security & Privacy

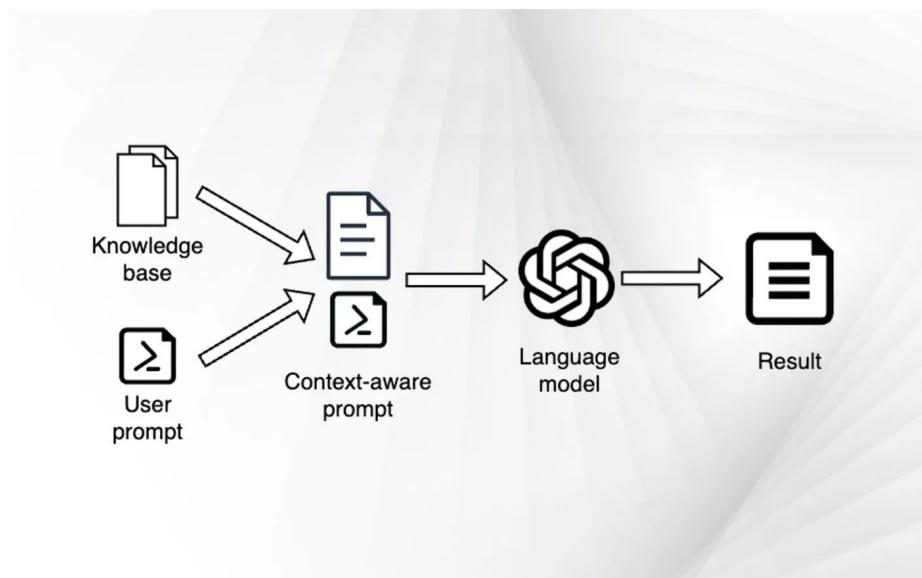


Data (Store) Security

Fine Tuning



LangChain Style Integrations

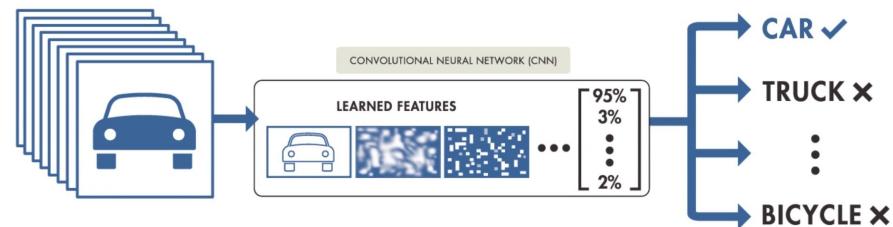


Data (Store) Security: Fine Tuning

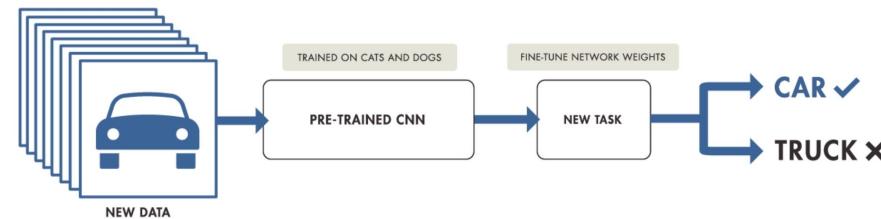
Areas for Concern

- Access Controls
- Data Poisoning Attacks
- Dataset Tampering

TRAINING FROM SCRATCH



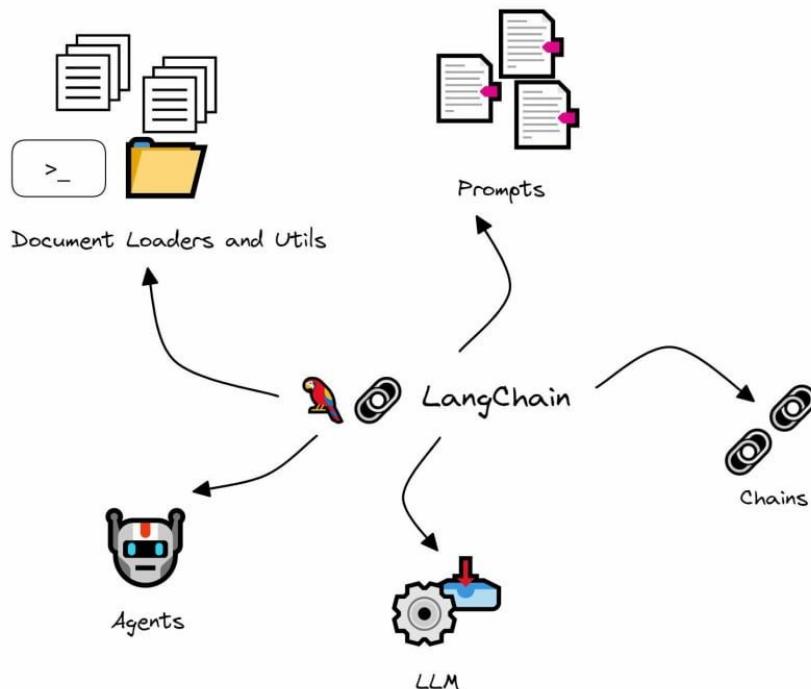
TRANSFER LEARNING



💡 Resource: [Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses](#)

Image Source: [Pinterest](#)

Data (Store) Security: LangChain Integration



Areas for Concern

- Access Controls
- Database Permissions
- API Security
- Code Vulnerabilities

LLM in the News



SECURITY / POLICY / TECH

ChatGPT's history bug may have also exposed payment info, says OpenAI



Photo by Amelia Holowaty Krales / The Verge

/ The company has published additional details and a technical breakdown of what lead it to take down the service on Monday.

By Mitchell Clark

Mar 24, 2023, 7:16 PM EDT | □ 4 Comments / 4 New



If you buy something from a Verge link, Vox Media may earn a commission. [See our ethics statement.](#)

Data Privacy & Regulatory Considerations

Considerations

- Sensitive Information Exposure
- Data Access Rights



Illustration by: MidJourney

Recommendations

- Data Anonymization
- Data Minimization

💡 Resource: [Privately Fine-Tuning Large Language Models with Differential Privacy](#)

💡 Resource: [ProPILE: Probing Privacy Leakage in Large Language Models](#)

Right to Be Forgotten in LLMs

Training of LLMs

- Training Data Memorizations
- Hallucinations



App Usage

- User Chat History
- In-Model Data



Resource: [Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions](#)

LLM in the News



Privacy

ChatGPT-maker OpenAI accused of string of data protection breaches in GDPR complaint filed by privacy researcher

Natasha Lomas @riptari / 1:01 PM EDT • August 30, 2023

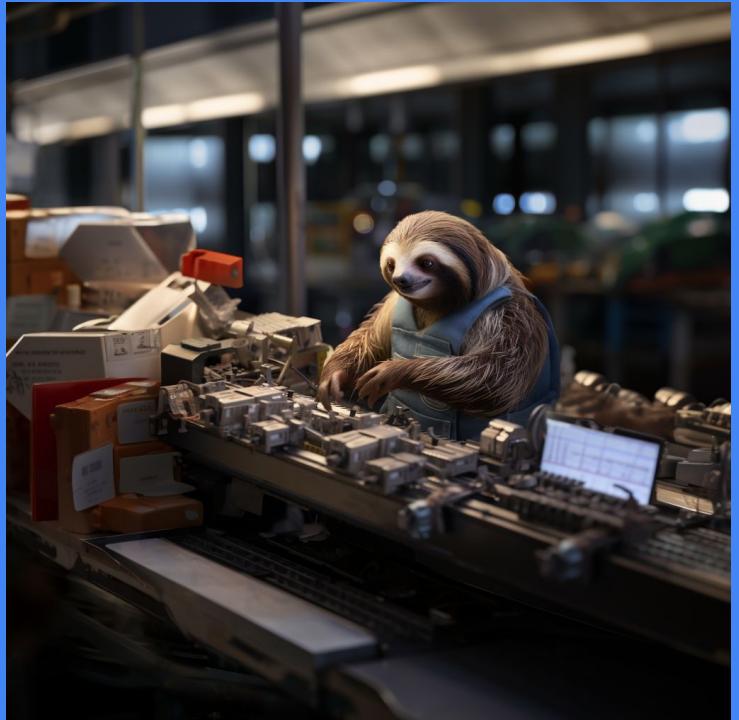
Comment



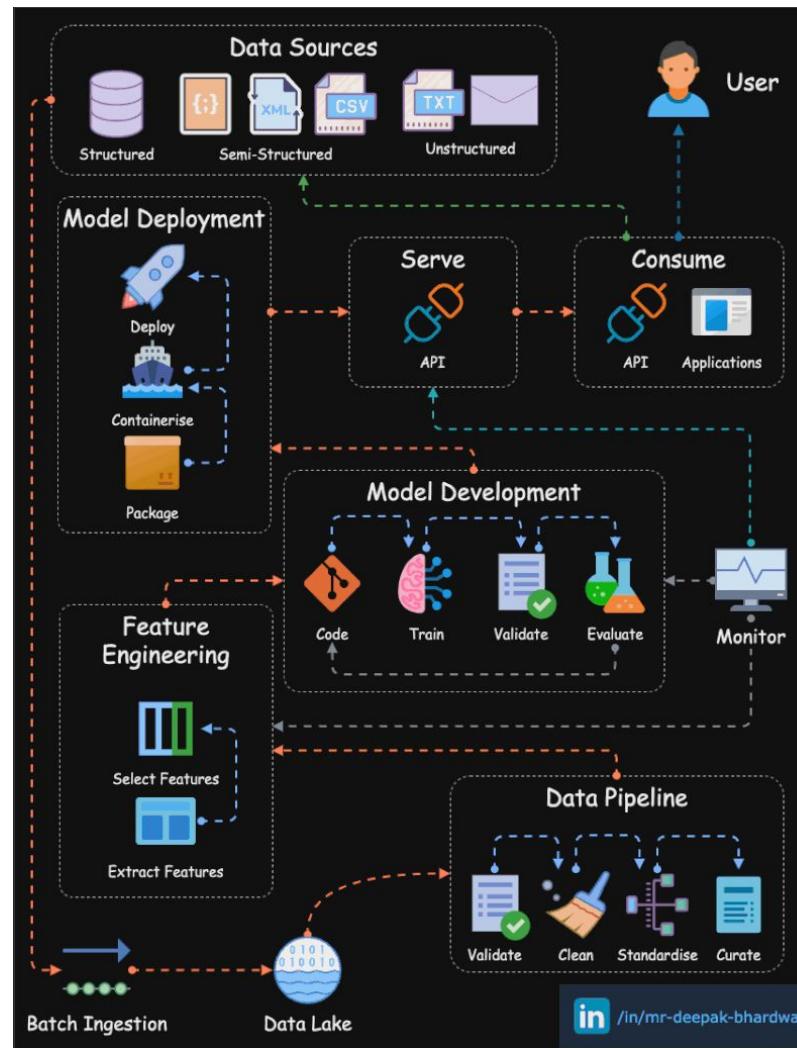
Image Credits: Leon Neal / Getty Images

💡 Resource: [ChatGPT-maker OpenAI accused of string of data protection breaches in GDPR complaint filed by privacy researcher](#)

ML Governance & Security



ML Pipeline Overview



ML Pipeline Overview



ML Model Governance Basics



Model Quality Control

- Measuring Performance Degradation
- Model Update Monitoring & Metrics



Resource: [Large Language Model Evaluation](#)

LLM in the News

ars TECHNICA

BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE STORE FORUMS

MODERN INDEMNIFICATION —

Microsoft offers legal protection for AI copyright infringement challenges

"Some customers are concerned about the risk of IP infringement claims," says Microsoft.

BENJ EDWARDS - 9/8/2023, 6:40 PM

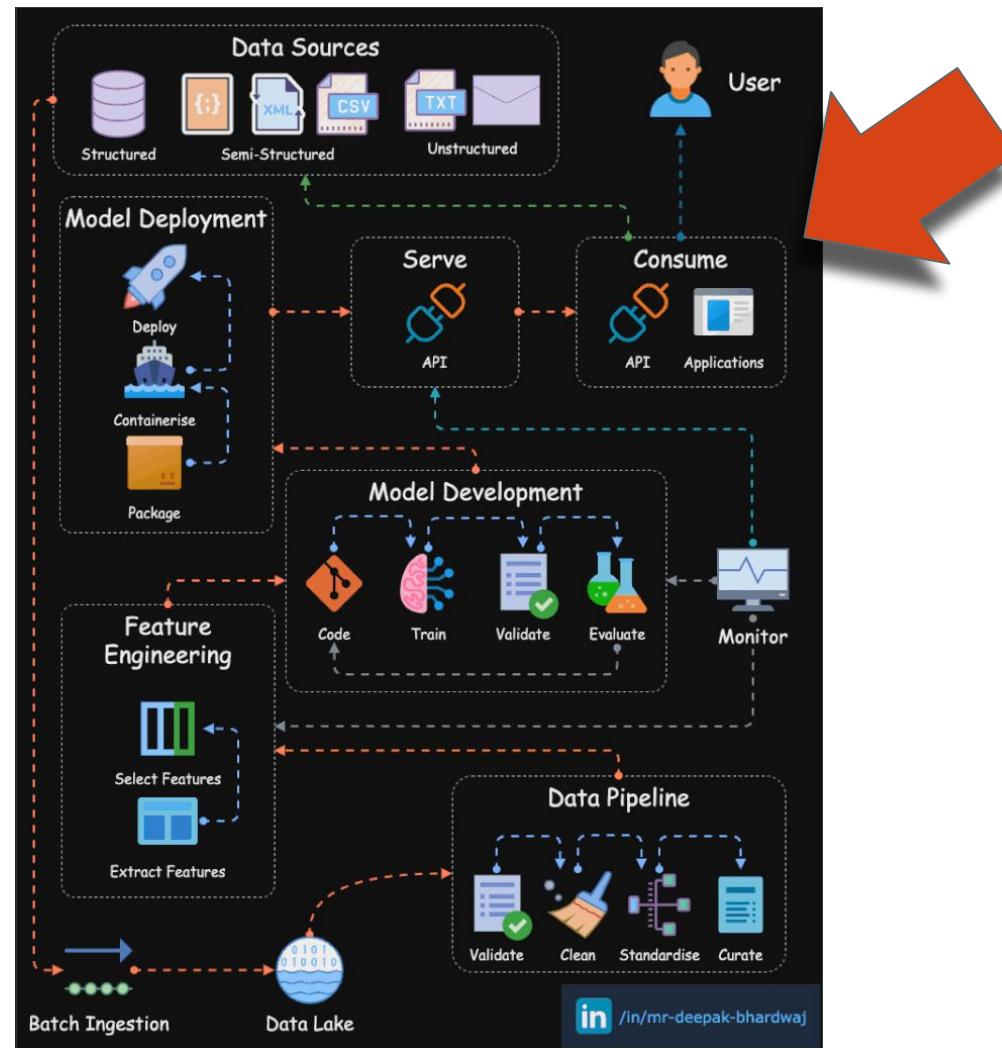


Getty Images / Benj Edwards

Enlarge

💡 Resource: [Microsoft offers legal protection for AI copyright infringement challenges](#)

ML Pipeline Overview



ML Pipeline Security: The Last Mile



- Prediction Service
 - Code Injection
 - API Vulnerability
 - DDoS
- General Application Security
 - Broken Access Controls
 - Security Misconfiguration
 - Authentication Failures

Securing ML Applications



ML Security Knowledge Bases



OWASP | OWASP Top 10 for LLM Applications v1.0.1

OWASP Top 10 for LLM Applications

LLM01

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02

Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03

Training Data Poisoning

Training data poisoning refers to manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors or biases that could compromise the model's security, effectiveness or ethical behavior.

LLM04

Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plugins add vulnerabilities.

LLM06

Sensitive Information Disclosure

LLMs may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. Implement data sanitization and strict user policies to mitigate this.

LLM07

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control due to lack of application control. Attackers can exploit these vulnerabilities, resulting in severe consequences like remote code execution.

LLM08

Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. This issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09

Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10

Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.



Resource: [OWASP Top 10 for LLM Applications](#)

ML Security Knowledge Bases

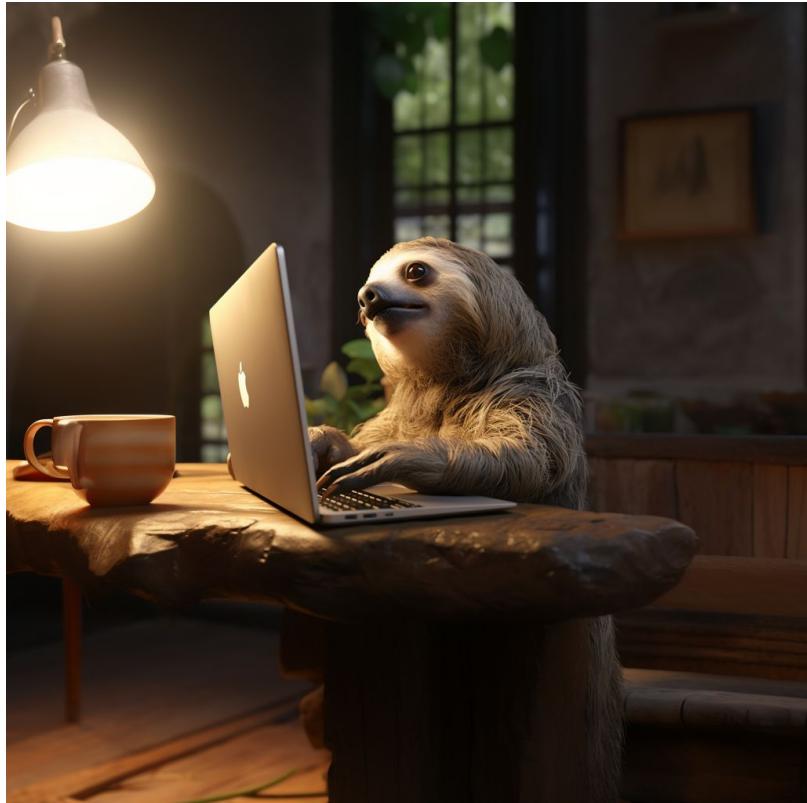
Mitre ATLAS

Reconnaissance &	Resource Development &	Initial Access &	ML Model Access	Execution &	Persistence &	Defense Evasion &	Discovery &	Collection &	ML Attack Staging	Exfiltration &	Impact &
5 techniques	7 techniques	4 techniques	4 techniques	2 techniques	2 techniques	1 technique	3 techniques	3 techniques	4 techniques	2 techniques	7 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	Evade ML Model	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Adversarial ML Attack Capabilities	Evade ML Model	Physical Environment Access				Discover ML Artifacts	Data from Local System &	Verify Attack		Spamming ML System with Chaff Data
Search Application Repositories	Exploit Public-Facing Application &	Full ML Model Access							Craft Adversarial Data		Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets										Cost Harvesting
	Poison Training Data										ML Intellectual Property Theft
	Establish Accounts &										System Misuse for External Effect

💡 Resource: [MITRE Atlas Matrix](#)

Parting Thoughts

- Data Security & Privacy
 - Data Strategy & Data Governance are needed
 - Same rules of data store security apply
 - Data Privacy can be more challenging with LLMs
- ML Pipeline Security Areas of Focus
 - Model Development & Deployment
 - Model Governance
 - Model Serving
 - General Application Security



Questions



hey@dafnz.com



@danielfernandez



/in/dafdz



danielfernandez@infosec.exchange



danielfernandez.medium.com



Presentation Slides: <https://github.com/dnlfdz/talks>