ACSAC 2023

Security for AI Real World LLMs
A FinSec Fine-Tuning Case Study

Dennis Moreau, PhD.
Sr. Director of Security Strategy – AI
joseph.dennis.moreau@intel.com
dennis.r.moreau@gmail.com

intel®

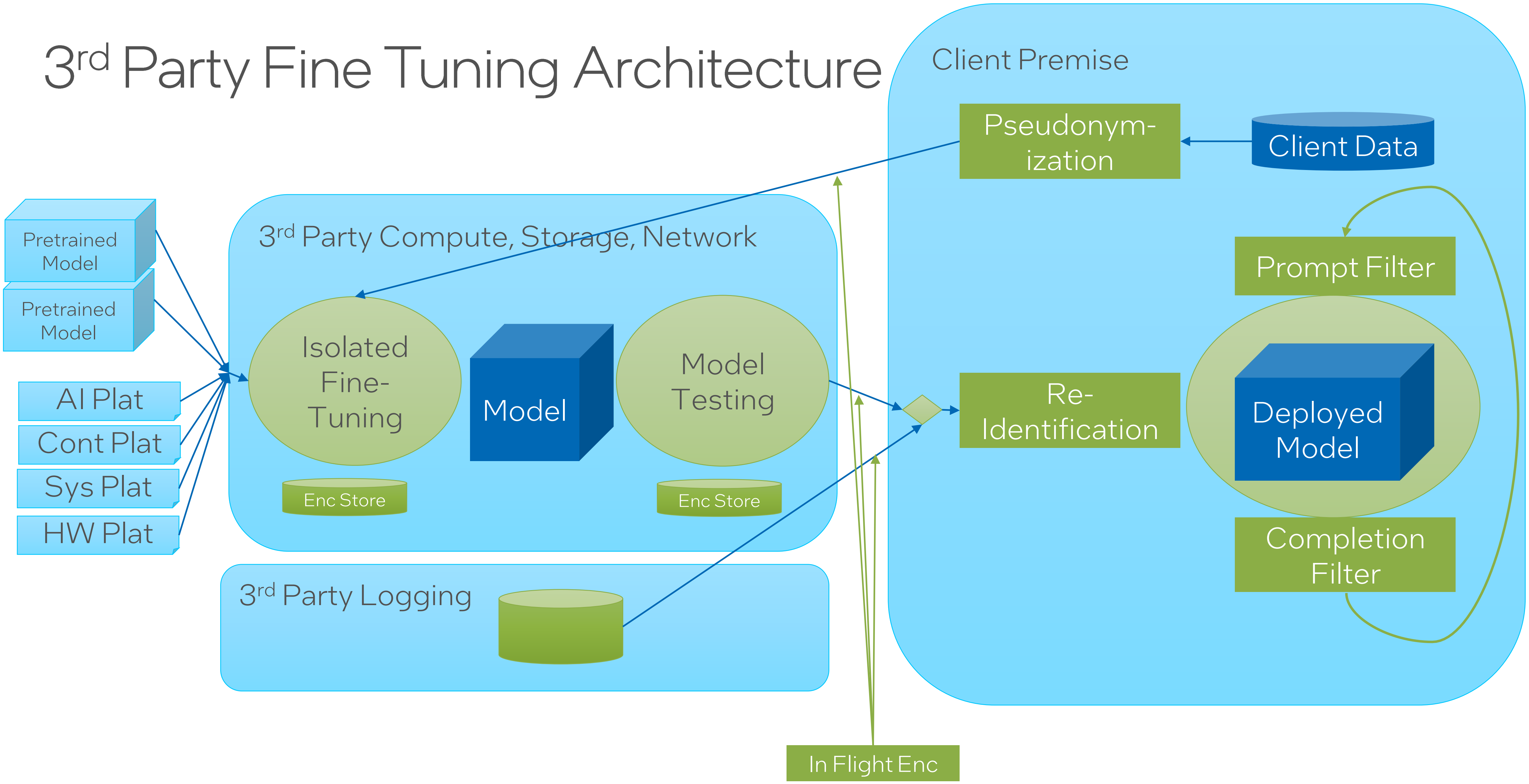
Case Study Scope

- 3rd Party Fine-Tuning of an LLM Model for a global financial sector consultancy
- Training data included regulated privacy information (GDPR) and corporate intellectual property (Client InfoSec).
- All fine tuning information was corporate owned.
- All intended usage was corporate internal
- 1st Order Threats:
  - Privacy leakage
  - Loss of critical intellectual property

Core Customer Security Requirements

- All Provided Corporate Training Data
  - Pseudonymized to GDPR standards
  - Encrypted at rest and in transit
- All 3rd Party Model Training Must Be Isolated
  - Compute, Network and Storage isolation.
  - Physical, Infrastructure and Temporal isolation.
- All privileged access must be logged and retained
- All storage scrubbed prior to, and after fine-tuning
- All artifacts shredded post engagement (data, model, intermediate artifacts)
- All supply chain items curated
- All software 3rd party scanned
- All system behavior logged
- Model verified – model performance, model privacy leakage

3rd Party Fine Tuning Architecture

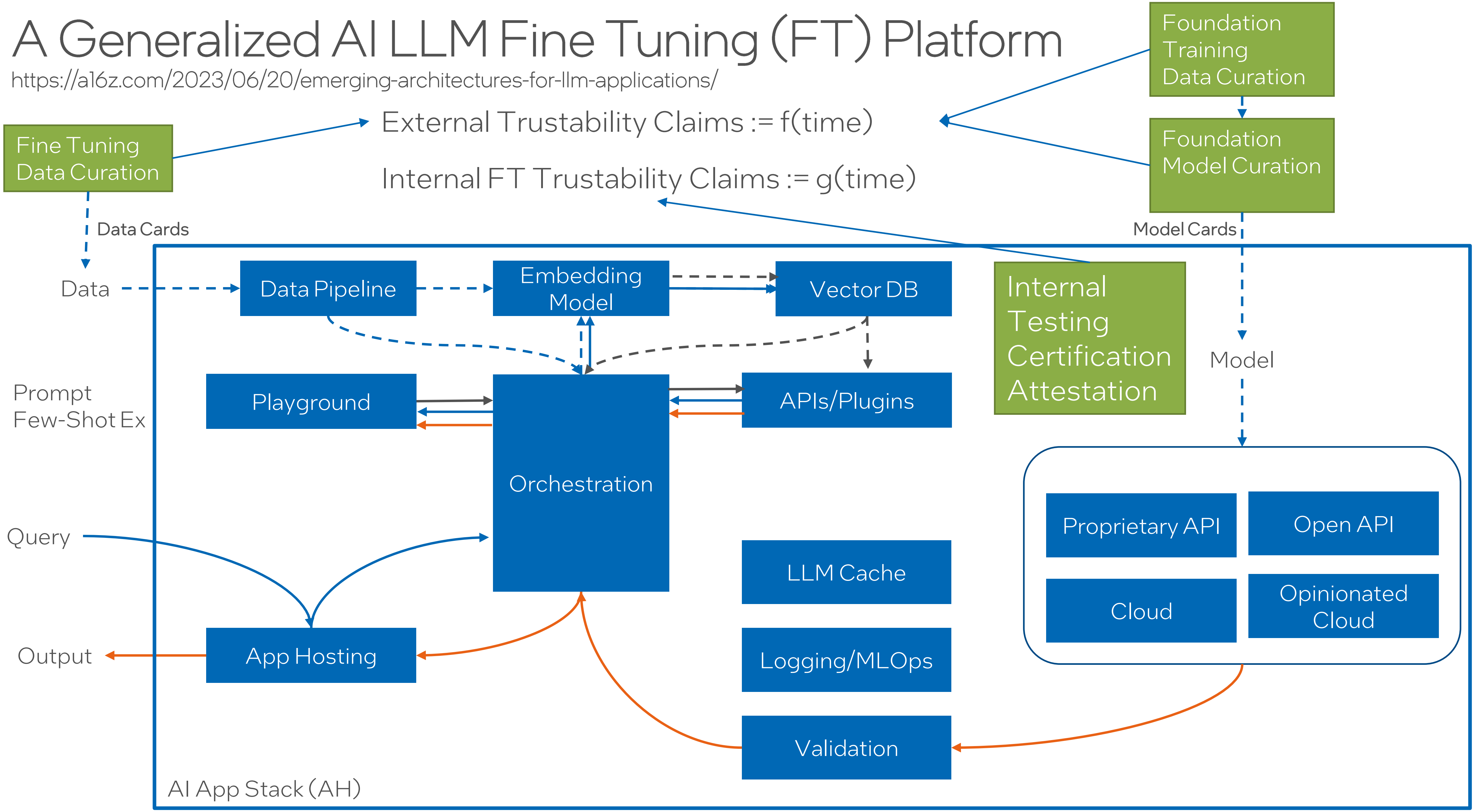


AI LLM & GenAI Cybersecurity is Different

Differences from conventional cybersecurity

A Generalized AI LLM Fine Tuning (FT) Platform

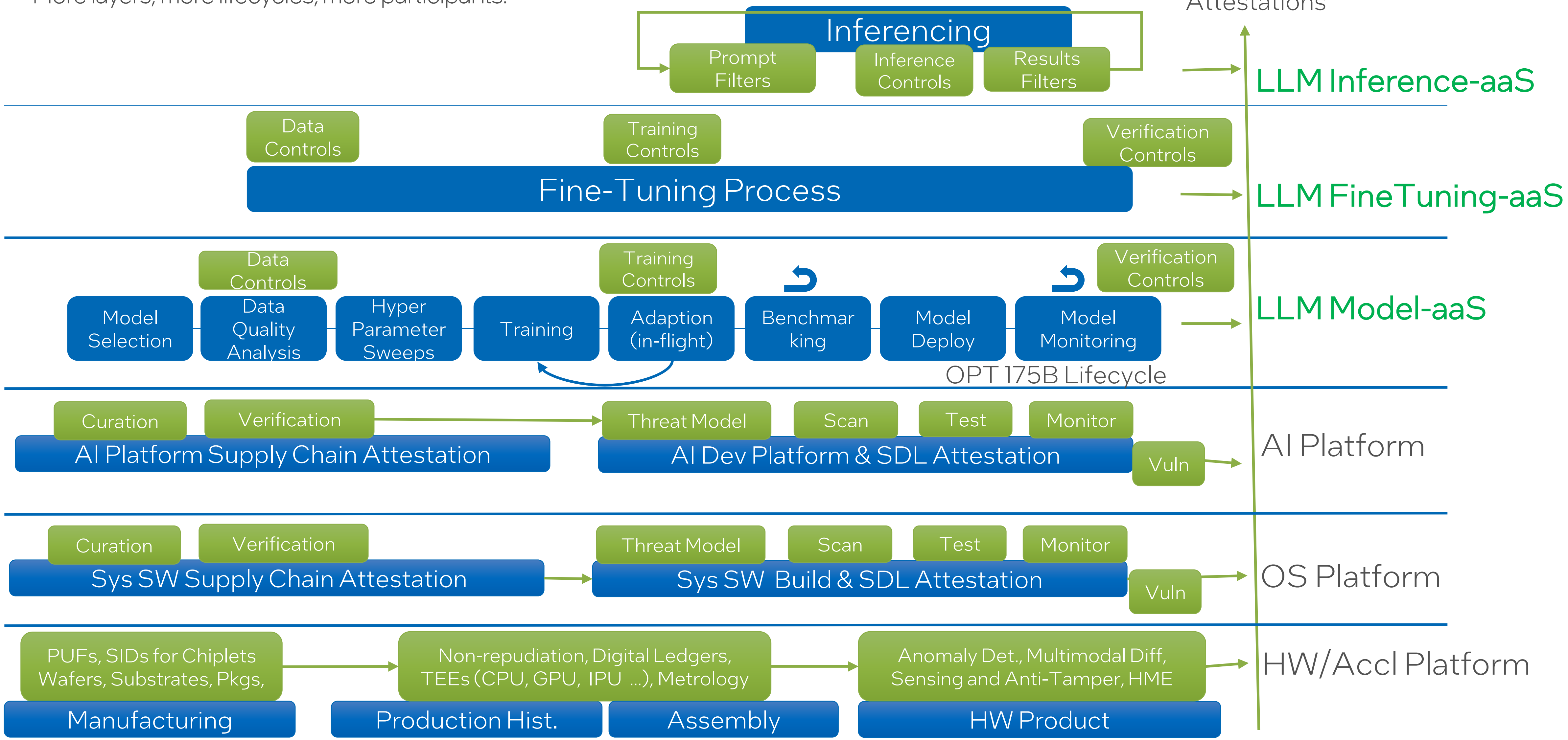
https://a16z.com/2023/06/20/emerging-architectures-for-llm-applications/



Andreesen Horowitz

What's new for cybersecurity?

More layers, more lifecycles, more participants.

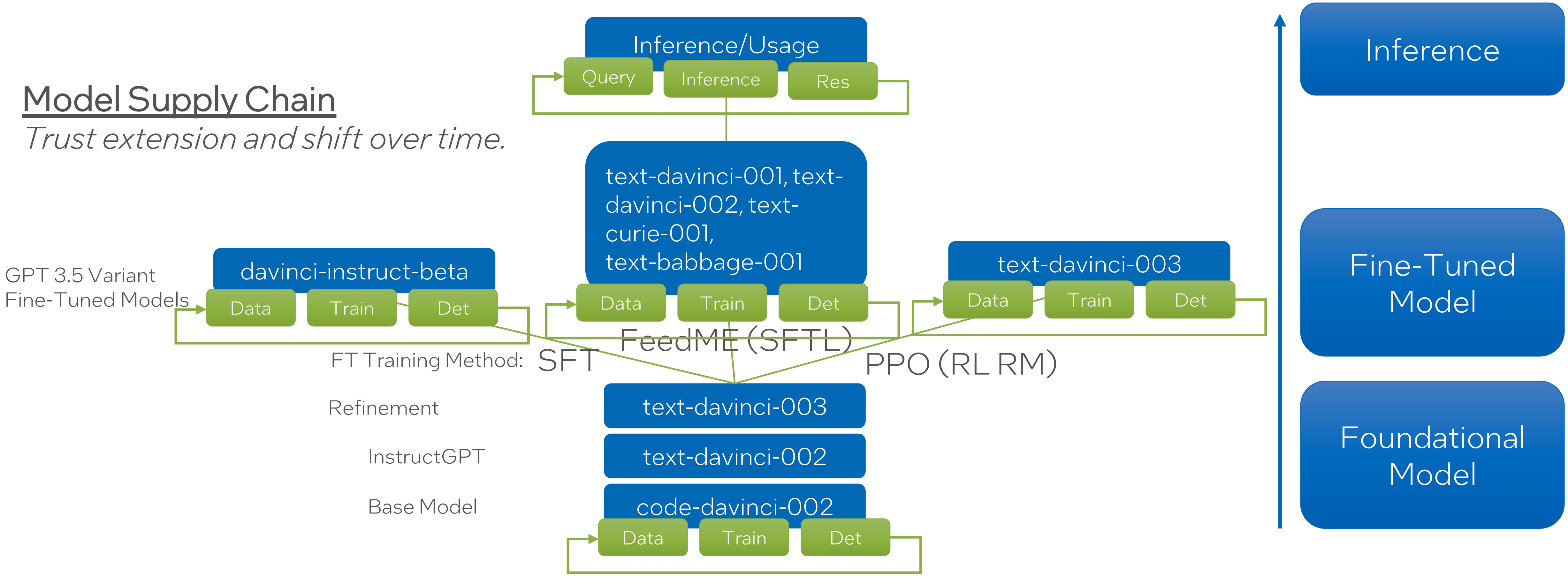


https://platform.openai.com/docs/model-index-for-researchers

Model Lifecycle Dependency

<u>Model Supply Chain</u>

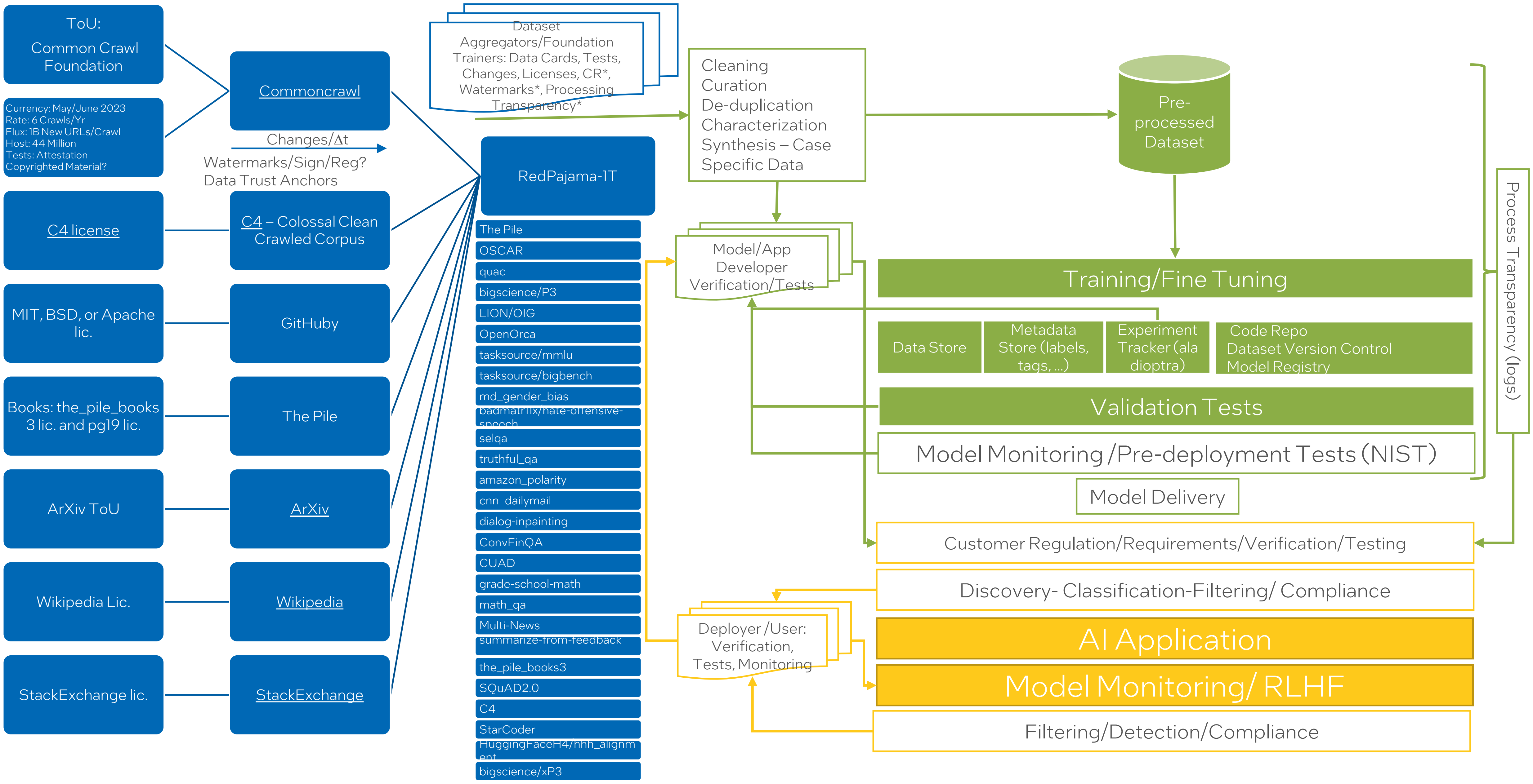
Trust extension and shift over time.



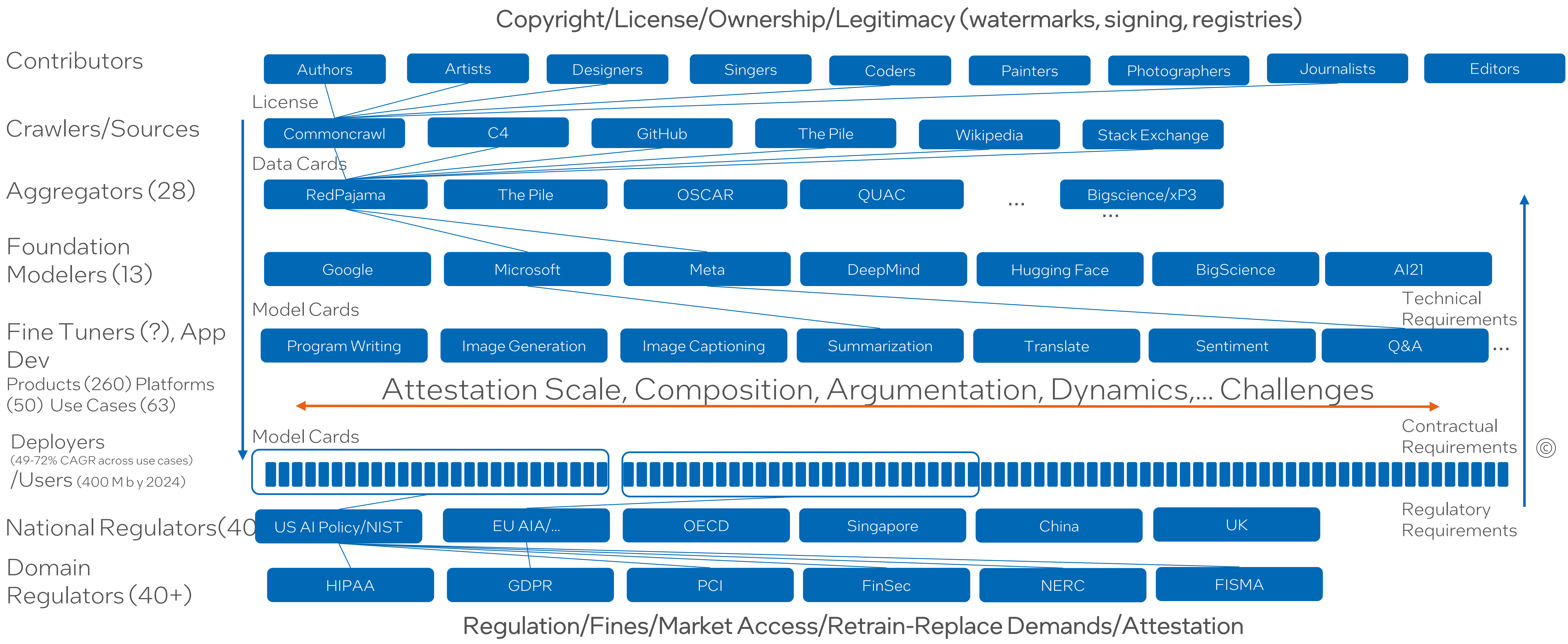
https://platform.openai.com/docs/model-index-for-researchers

Data Source Lifecycle Dependencies

Applies to all training data characterizations, preparation, controls and adaptation based on observation, tests and requirements



Trust (attestation) Dependencies
Complex and Dynamic

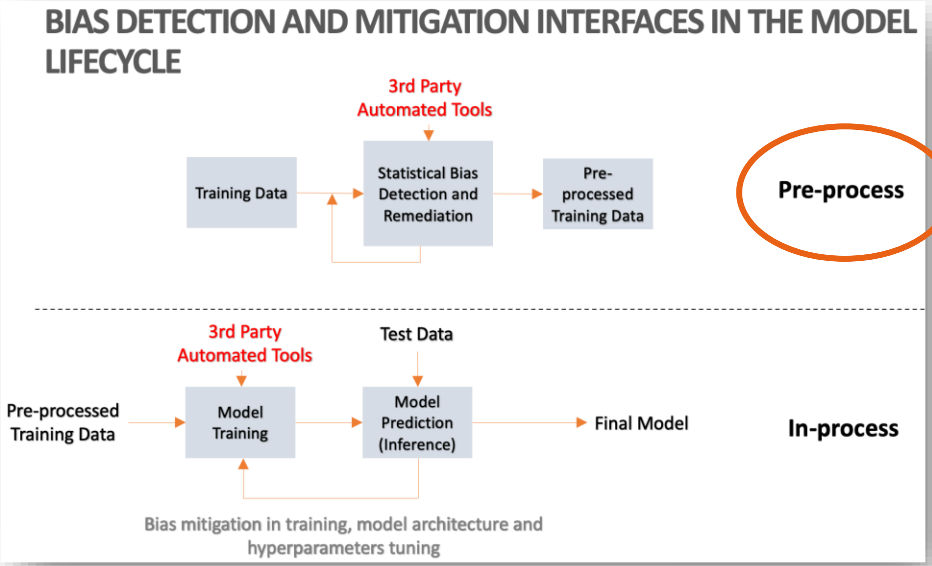


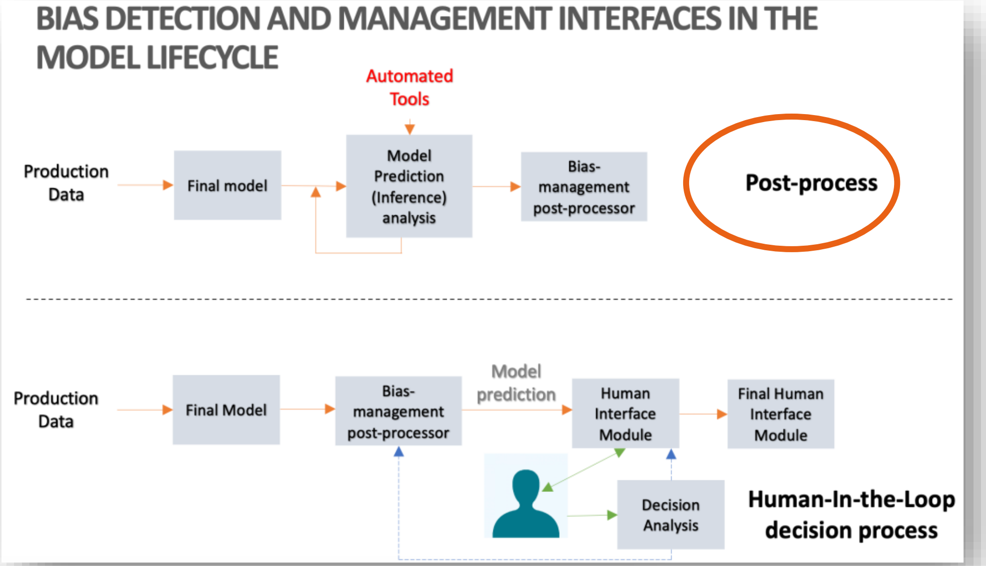
AI Cybersecurity – Future Requirements

AI Security Frameworks/Architectures: A System of Controls

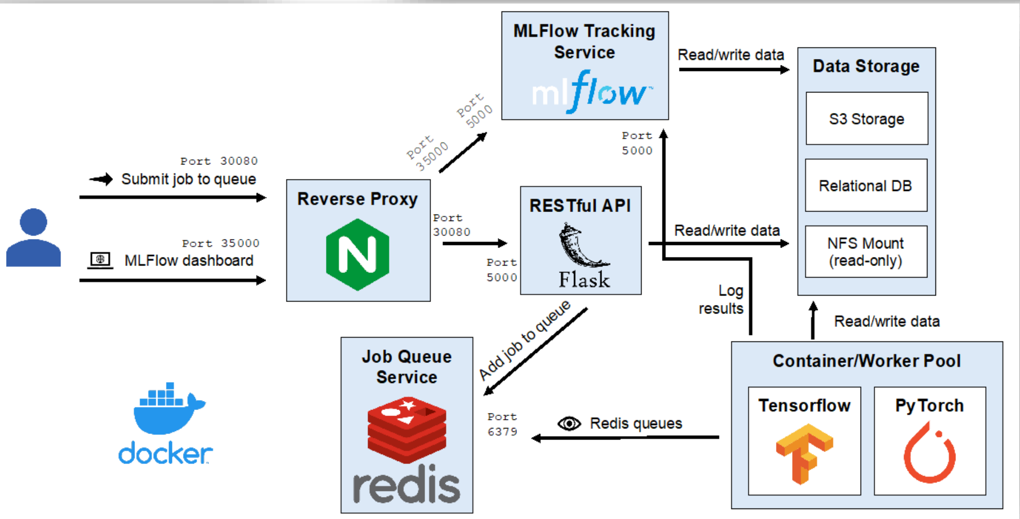
Example: Bias Control tests and evidence exist across the AI Lifecycle

(From: MITIGATING AI/ML BIAS IN CONTEXT: Establishing Practices for Testing, Evaluation, Verification, and Validation of AI Systems - https://www.nccoe.nist.gov/sites/default/files/2022-11/ai-bias-pd-final.pdf Adaptation to GenAI and LLMs underway in GAI-WG, see slide 1)

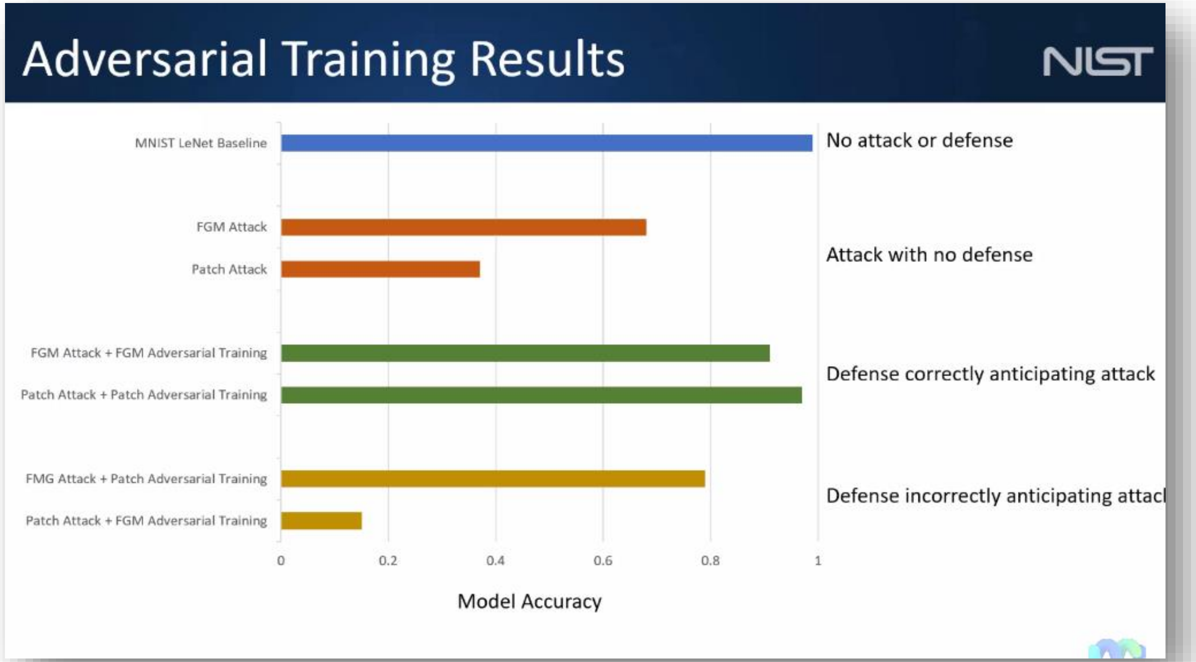




NIST DIOPTRA Test Framework



Model Testing: Interference (example)



NIST Dioptra Observation

The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Conflicting Interactions among Protection Mechanisms for Machine Learning Models

Sebastian Szyller[1], N. Asokan [2,1]
[1]Aalto University
[2]University of Waterloo
contact@sebszyller.com, asokan@acm.org

Abstract

Nowadays, systems based on machine learning (ML) are widely used in different domains. Given their popularity, ML models have become targets for various attacks. As a result, research at the intersection of security/privacy and ML has flourished. Typically such work has focused on individual types of security/privacy concerns and mitigations thereof.

However, in real-life deployments, an ML model will need to be *protected against several concerns simultaneously*. A *protection mechanism* optimal for a specific security or privacy concern may interact negatively with mechanisms intended to address other concerns. Despite its practical relevance, the potential for such conflicts has not been studied adequately.

In this work, we first provide a framework for analyzing such *conflicting interactions*. We then focus on systematically analyzing pairwise interactions between protection mechanisms for one concern, *model and data ownership verification*, with two other classes of ML protection mechanisms: *differentially private training*, and *robustness against model evasion*. We find that several pairwise interactions result in conflicts.

We also explore potential approaches for avoiding such conflicts. First, we study the effect of hyperparameter relaxations, finding that there is no sweet spot balancing the performance of both protection mechanisms. Second, we explore whether modifying one type of protection mechanism (ownership verification) so as to decouple it from factors that may be impacted by a conflicting mechanism (differentially private training or robustness to model evasion) can avoid conflict. We show that this approach can indeed avoid the conflict between ownership verification mechanisms when combined with differentially private training, but has no effect on robustness to model evasion. We conclude by identifying the gaps in the landscape of studying interactions between other types of ML protection mechanisms.

1 Introduction

Machine learning (ML) models constitute valuable intellectual property. They are also increasingly deployed in risk-sensitive domains. As a result, various security and privacy requirements for ML model deployment have become apparent. This, in turn, has led to substantial recent research at the intersection of machine learning and security/privacy. The research community largely focuses on individual types of security/privacy threats and ways to defend against them. This facilitates iterative improvements, and allows practitioners to evaluate the benefit of any new approaches.

In this work, we argue that in realistic deployment setting, multiple security/privacy concerns need to be considered *simultaneously*. Therefore, any *protection mechanism* for a particular concern, needs to be tested together with defences against *other common concerns*. We show that when deployed together, ML protection mechanisms may not work as intended due to *conflicting interactions* among them.

We claim the following contributions:

1) We highlight the importance of understanding *conflicting interactions* among ML protection mechanisms, and provide a framework for studying it (Section 3).
2) We use our framework to analyse the interaction between *model ownership verification mechanisms* with two other types of protection mechanisms: *differentially private training* and *adversarial training*. We provide a theoretical justification (Section 4) for each potential pairwise conflict, and evaluate it empirically (Sections 5 and 6).
3) We explore whether conflicts can be avoided by changing (a) the hyperparameters of each protection mechanism, or (b) the design of the mechanism itself (Section 7).

2 Background

2.1 Machine Learning

The goal of a ML classification model $F_{\mathcal{V}}$ trained on some dataset $\mathcal{D}_{\text{TR}}$ is to perform well on the given classification task according to some metric ϕ measured on a test set $\mathcal{D}_{\text{TE}}$. The whole dataset is denoted as $\mathcal{D} = \{\mathcal{D}_{\text{TR}}, \mathcal{D}_{\text{TE}}\}$. An individual record consists of an input x and the corresponding label y. Throughout this work, we use the accuracy metric $\phi_{\text{ACC}}(F_{\mathcal{V}}, \mathcal{D}_{\text{TE}})$ to assess a model $F_{\mathcal{V}}$ using $\mathcal{D}_{\text{TE}}$:

$$\phi_{\text{ACC}}(F_{\mathcal{V}}, \mathcal{D}_{\text{TE}}) = \frac{1}{|\mathcal{D}_{\text{TE}}|} \sum_{x \in \mathcal{D}_{\text{TE}}} \mathbb{1}(\hat{F}_{\mathcal{V}}(x) = y). \quad (1)$$

where $F_{\mathcal{V}}(x)$ is the full probability vector and $\hat{F}_{\mathcal{V}}(x)$ is the most likely class.

2.2 Ownership Verification

Emerging understanding of LLM trust-ability limits

Universal and Transferable Adversarial Attacks on Aligned Language Models

Andy Zou[1], Zifan Wang[2], J. Zico Kolter[1,3], Matt Fredrikson[1]
[1]Carnegie Mellon University, [2]Center for AI Safety, [3]Bosch Center for AI
andyzou@cmu.edu, zifan@safe.ai, zkolter@cs.cmu.edu, mfredrik@cs.cmu.edu

arXiv:2307.15043v1 [cs.CL] 27 Jul 2023

Because "out-of-the-box" lar
deal of objectionable content, re
attempt to prevent undesirable g
cumventing these measures—so-
required significant human ingen
adversarial prompt generation h
propose a simple and effective a
generate objectionable behaviors
attached to a wide range of queri
to maximize the probability tha
than refusing to answer). Howev
proach automatically produces t
and gradient-based search techni
generation methods.
Surprisingly, we find that the
quite *transferable*, including to bl
an adversarial attack suffix on *m*
types of objectionable content), a
13B). When doing so, *the resu*
able content in the public int

A LLM Assisted Exploitation of AI-Guardian

Nicholas Carlini
Google DeepMind

v:2307.15008v1 [cs.CR] 20 Jul 2023

Abstract

Large language models (LLMs) are now highl
at a diverse range of tasks. This paper studie
or not GPT-4, one such LLM, is capable of as
searchers in the field of adversarial machine lear
case study, we evaluate the robustness of AI-Gu
recent defense to adversarial examples publishe
S&P 2023, a top computer security conference.
pletely break this defense: the proposed scheme
increase robustness compared to an undefended
We write none of the code to attack this mod
stead prompt GPT-4 to implement all attack a
following our instructions and guidance. Th
was surprisingly effective and efficient, with the
model at times producing code from ambiguou
tions faster than the author of this paper could
We conclude by discussing (1) the warning sig
in the evaluation that suggested to us AI-Guard
be broken, and (2) our experience with designi
and performing novel research using the most
vances in language modeling.

1 Introduction

Defending against adversarial examples is hard
ically, the vast majority of adversarial exampl

Removing RLHF Protections in GPT-4 via Fine-Tuning

Qiusi Zhan[1], Richard Fang[1], Rohan Bindu[1], Akul Gupta[1],
Tatsunori Hashimoto[2], Daniel Kang[1]

[1]UIUC, [2]Stanford University

Nov 2023

Abstract

As large language models (LLMs) have increased in their capabilities, so does their potential for dual use. To reduce harmful outputs, produces and vendors of LLMs have used reinforcement learning with human feedback (RLHF). In tandem, LLM vendors have been increasingly enabling fine-tuning of their most powerful models. However, concurrent work

RLHF can be a powerful method to reduce harmful outputs.

However, these API providers are increasingly providing methods to fine-tune the API-gated models, such as GPT-4. Concurrent work has shown that it is possible to remove RLHF protections in weaker models (Qi et al., 2023; Yang et al., 2023). This raises an important question: can we use fine-tuning to remove RLHF protections in state-of-the-

Whitehouse AI Executive Order

Safe, Secure and Trustworthy AI

- <date>
- AI EO 2023 exhibits a few main policy objectives
- Each objective has delegated actions that may include analysis, policy, planning,, guidance and programmatic efforts.

Whitehouse AI EO: Timeline Part 1

Date		
Current	Federal Trade Commission	Consider use of rulemaking and regulatory authority, for fair competition in AI and to protect consumers from unfair and deceptive practices
	Sec of Labor	Consider use of authority to protect users from fraud, discrimination, privacy threats, and emergent risks , from the us of AI
	Sec of Labor	Clarification of monitoring and transparency requirements for third party AI services, and employment of AI by independent agencies.
01-28-2024	Sec of Commerce	Reporting requirements for dual-use foundational models and computing clusters
	Dept of HHS	Establish HHS AI Task Force
02-27-2024	US PTO Director	Patent Examiner and Applicant guidance on IP, Inventorship and the use of AI
03-28-2024	Sec of the Treasury	Report on best practices to manage AI-specific Cybersecurity Risks for financial Institutions
(cont'd)		

Whitehouse AI EO: Timeline Part 2

Date		
04-27,2024	Sec of Commerce	Recommended regulations requiring foreign resellers of IaaS potential AI training capacity, to identify foreign users
	Sec of Labor	Publish Best practices for employers to mitigate harm and maximize benefits to employees, of AI
	Sec of Homeland Sec Sec of Commerce	Inclusion of AI safety and security guidance into CIS operator guidelines
06-26-2024	Sec of Commerce	Report on existing methods and development of methods for detecting, labeling and limiting/preventing AI generated content
07-26-2024	Sec of Commerce, Energy & Homeland Security	Guidelines for developing safe, secure trustworthy AI and validation (RT, Testing, …)
	Asst to the Pres Nat Sec Asst to the Pres and DCoS for Policy	National Security Memorandum on AI
	USPTO Director Director of US Copyright Office	Recommendations on EOs related to Copyright and AI
	Sec of Commerce Sec of State	Report on risks and benefits of widely-available dual-use foundational models
10-29-2024	Sec of Labor	Publish Fed Contractor Guidance on non-discrimination in AI and automated hiring
> 12-23-2024	FAR Council	Amend FARs to align on labeling and authenticating published content

But not all AI Regulation Efforts are "Aligned"

Version 1 (2024)	EU	US
Structure	Comprehensive Unified Policy	By Sector
Objective	Risk Moderated Regulatory FW	Benefits vs Risk Balance
Critical AI	Certification for High Risk AI	Safety, Trust, Responsibility for all AI
Innovation Impact	Universality objective may impede innovation	Flexibility intended to accommodate innovation
Participation	By member nation – Parliament	By agency with industry and public WGs
Schedule	2024- Enactment 2026- Implementation	Recommendations/analyses due by Nov 2024

Conclusion: Case Study Results /Proof Points

- The PoC for 3rd Party FT of LLM Models in regulated FinSec domain
  - Satisfied client regulatory requirements – GDPR deployment and production
  - Satisfied client risk tolerance – Information protection, Model protection,
  - Ref Architecture achieved
    - Shared accelerator tolerant multi-tenancy (sequential tenancy)
    - Network, compute, storage and temporal isolation
    - Comprehensive observability
- The Platform architecture and software were successfully productized
- Established a foundation for addressing emerging AI regulatory requirements
- Caveat: We are no where near the AI Cybersecurity or AI Trustability finish-line

Q&A

Dennis Moreau

Sr. Director Security Strategy for AI

DCAI

joseph.dennis.moreau@intel.com

719-964-0836

intel®