HPC + AI WALL STREET

2024

Responsible AI

Industry Trends and Key Considerations for AI Practitioners

HPC | QUANTUM | DATA | AI

Disclaimer

- The views expressed by the presenters are not necessarily those of Ernst & Young LLP or other members of the global EY organization.
- These slides are for educational purposes only and are not intended to be relied upon as accounting, tax or other professional advice. Please refer to your advisors for specific advice.
- Neither EY nor any member firm thereof shall bear any responsibility whatsoever for the content, accuracy, or security of any third-party websites that are linked (by way of hyperlink or otherwise) in this presentation.

Speakers

Rani Bhuva

Principal, EY Americas Financial Services Responsible AI Leader
Ernst & Young LLP (US)

Kiranjot Dhillon

Senior Manager, EY Americas Financial Services AI Leader
Ernst & Young LLP (US)

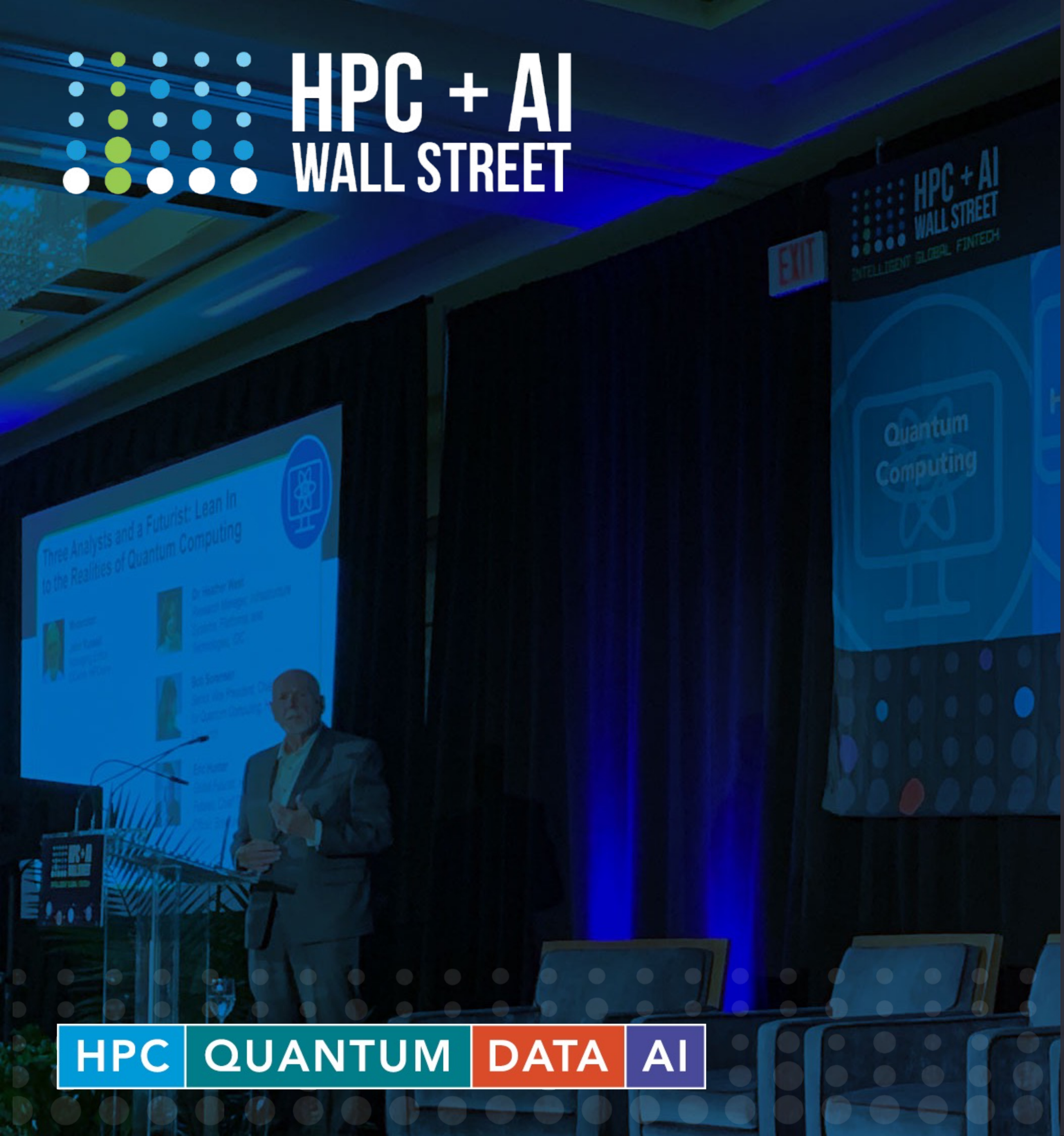

HPC + AI
WALL STREET
EXIT
HPC + AI WALL STREET
INTELLIGENT GLOBAL FINTECH
Quantum Computing
Three Analysts and a Futurist: Lean In to the Realities of Quantum Computing
HPC QUANTUM DATA AI

Table of Contents

01 AI — regulatory landscape

02 Emerging compliance trends

03 AI development and operationalization

Q&A

Summary of AI Regulatory and Policy Developments

United States

The federal and state governments are regulating artificial intelligence (AI) through a series of instruments (executive orders, acts, risk management frameworks, etc.).

European Union

The EU has adopted comprehensive, risk-based legislation (the EU AI Act), which is expected to become effective gradually over 24 months, starting in Q4 ‘24/Q1 ‘25 with mandates on prohibited use cases.

United Kingdom

UK is following a self-proclaimed light-touch, pro-innovation approach to AI regulation, with a high-level AI bill introduced in Parliament.

Canada

Canada has issued comprehensive, risk-based legislation (the AI and Data Act), which is expected to become effective in 2025. The government is also making substantial investments to position Canada as an AI leader globally.

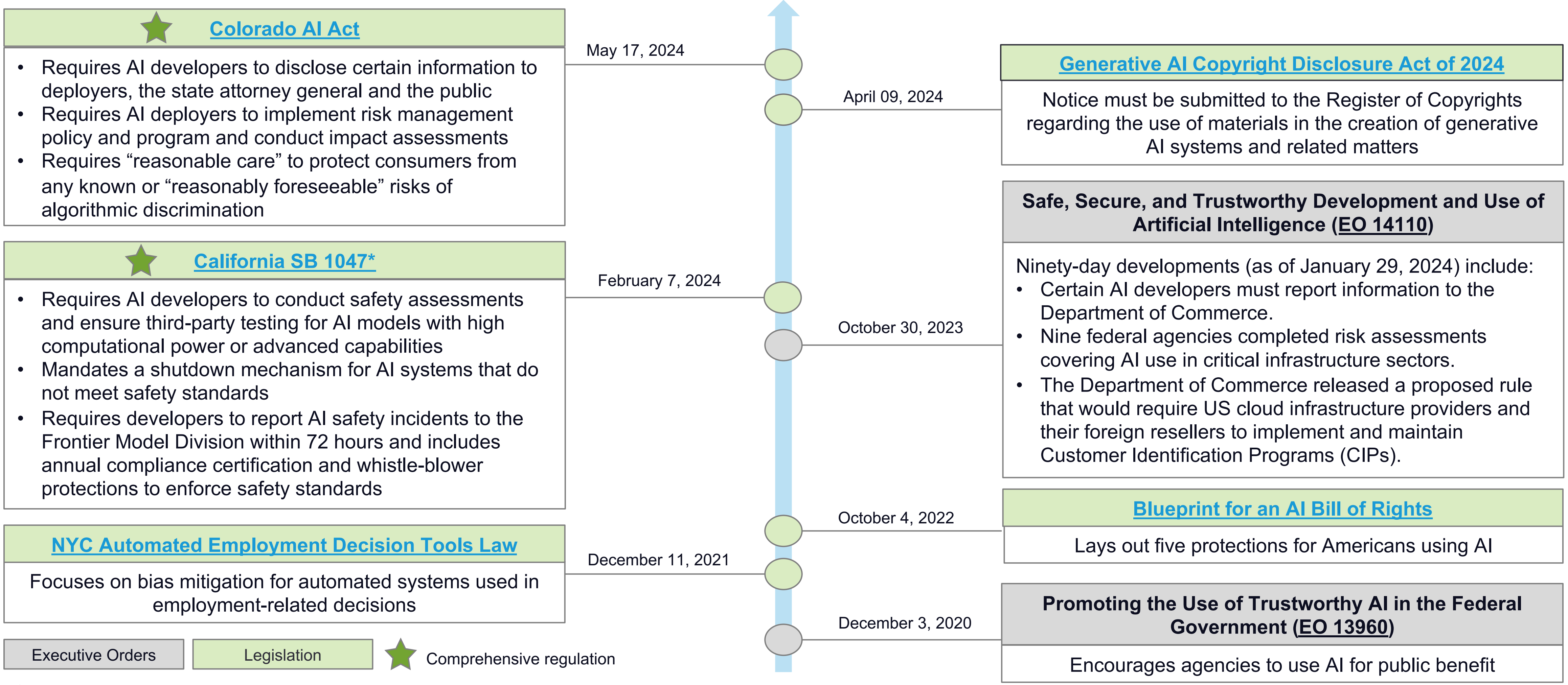
Singapore

Singapore has released its AI National Strategy 2.0., which is complemented with industry-specific, responsible AI principles (e.g., FEAT principles for FS).

G7

The G7 has issued the International Guiding Principles on Artificial intelligence and a voluntary Code of Conduct for AI developers.

Key US AI Regulatory and Policy Development



★ **Colorado AI Act**

May 17, 2024

- Requires AI developers to disclose certain information to deployers, the state attorney general and the public
- Requires AI deployers to implement risk management policy and program and conduct impact assessments
- Requires “reasonable care” to protect consumers from any known or “reasonably foreseeable” risks of algorithmic discrimination

Generative AI Copyright Disclosure Act of 2024

April 09, 2024

Notice must be submitted to the Register of Copyrights regarding the use of materials in the creation of generative AI systems and related matters

★ **California SB 1047***

February 7, 2024

- Requires AI developers to conduct safety assessments and ensure third-party testing for AI models with high computational power or advanced capabilities
- Mandates a shutdown mechanism for AI systems that do not meet safety standards
- Requires developers to report AI safety incidents to the Frontier Model Division within 72 hours and includes annual compliance certification and whistle-blower protections to enforce safety standards

Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (EO 14110)

October 30, 2023

Ninety-day developments (as of January 29, 2024) include:

- Certain AI developers must report information to the Department of Commerce.
- Nine federal agencies completed risk assessments covering AI use in critical infrastructure sectors.
- The Department of Commerce released a proposed rule that would require US cloud infrastructure providers and their foreign resellers to implement and maintain Customer Identification Programs (CIPs).

Blueprint for an AI Bill of Rights

October 4, 2022

Lays out five protections for Americans using AI

NYC Automated Employment Decision Tools Law

December 11, 2021

Focuses on bias mitigation for automated systems used in employment-related decisions

Promoting the Use of Trustworthy AI in the Federal Government (EO 13960)

December 3, 2020

Encourages agencies to use AI for public benefit

Executive Orders | Legislation | ★ Comprehensive regulation

*Yet to be enacted

AI Regulatory Trends and Expectations for US Banks

The How

Leading practices

Global jurisdictions

Professional organizations

Sector & use-case specific

US national regulatory actions

USTD	CFPB	FINRA
OCC	FHFA, HUD	SEC

Alignment with NIST

The What

AI governance framework

AI inventory

Reporting on AI-specific risks

Generative AI Risk Taxonomy Continues to Evolve

Heightened Risks for Generative AI (NIST AI 600-1)

Confabulation

Confidently stated but erroneous or false content

Data privacy

Leakage and unauthorized disclosure or de-anonymization of sensitive data

Information integrity

Generation, exchange and consumption of content that may not be vetted, accurate, etc.

Information security

GenAI-powered threats and increased available attack surface for targeted cyber attacks

Intellectual property

GenAI may produce and disseminate copyrighted or trademarked content, or provide information that enables patent infringement

Value chain

Non-transparent or untraceable integration of upstream components across the AI lifecycle

Environmental

INegative ecosystem impact due to high resource usage in training models

Toxicity, bias and homogenization

GenAI models may produce toxic or biased content due to inherent biases in training data or feedback loops

Human-AI configuration

Arrangement or interaction of humans and AI systems that can result in algorithmic aversion, automation bias, overreliance, anthropomorphization, etc.

Obscene, degrading and abusive content

GenAI may intensify the spread of harmful and explicit content, heightening privacy/safety risks

Dangerous or violent recommendations

GenAI eases the ability for bad faith actors to generate violent, inciting or threatening content

CBRN information

Eased access to nefarious information related to chemical, biological, radiological or nuclear (CBRN) weapons

+

Risk Carried Over from Existing AI Models

Data capability

Existing data capabilities (e.g., data modeling, storage, processing) and data governance (e.g., lineage and traceability) may not be sufficient for fine-tuning and business use of GenAI

Technology capability

GenAI adoption increases the computational needs and therefore potentially impacts the current use of infrastructure by other business use

Security

Training data and trained GenAI model may be leaked out of the institution or vendor platform due to cyber attack or adversarial prompt engineering

Bias/fairness

Large volume of training data used in pre-training may introduce bias and unfairness

Complex model and training process make it hard to identify and control bias

Business continuity

Heavy reliance on third-party pretrained complex GAI, may aggravate the business continuity

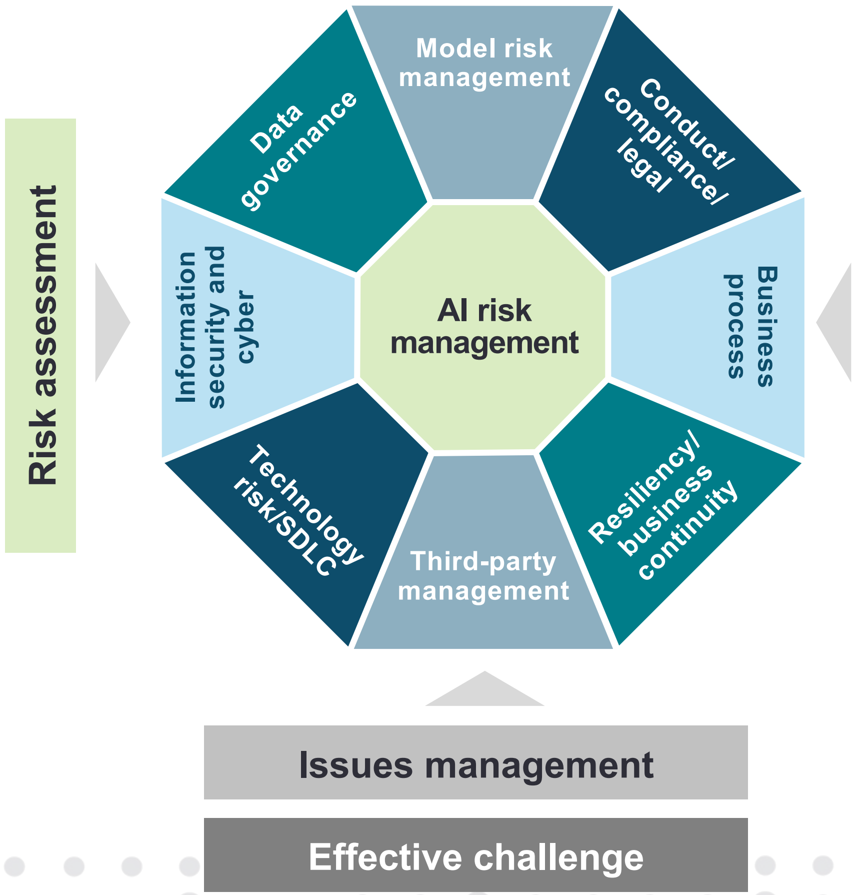
Explainability

Large models can make the GenAI a black box, which lacks explainability

Source: NIST AI 600-1 publication

Effective and Responsible AI Governance Requires Enterprise-wide Coordination

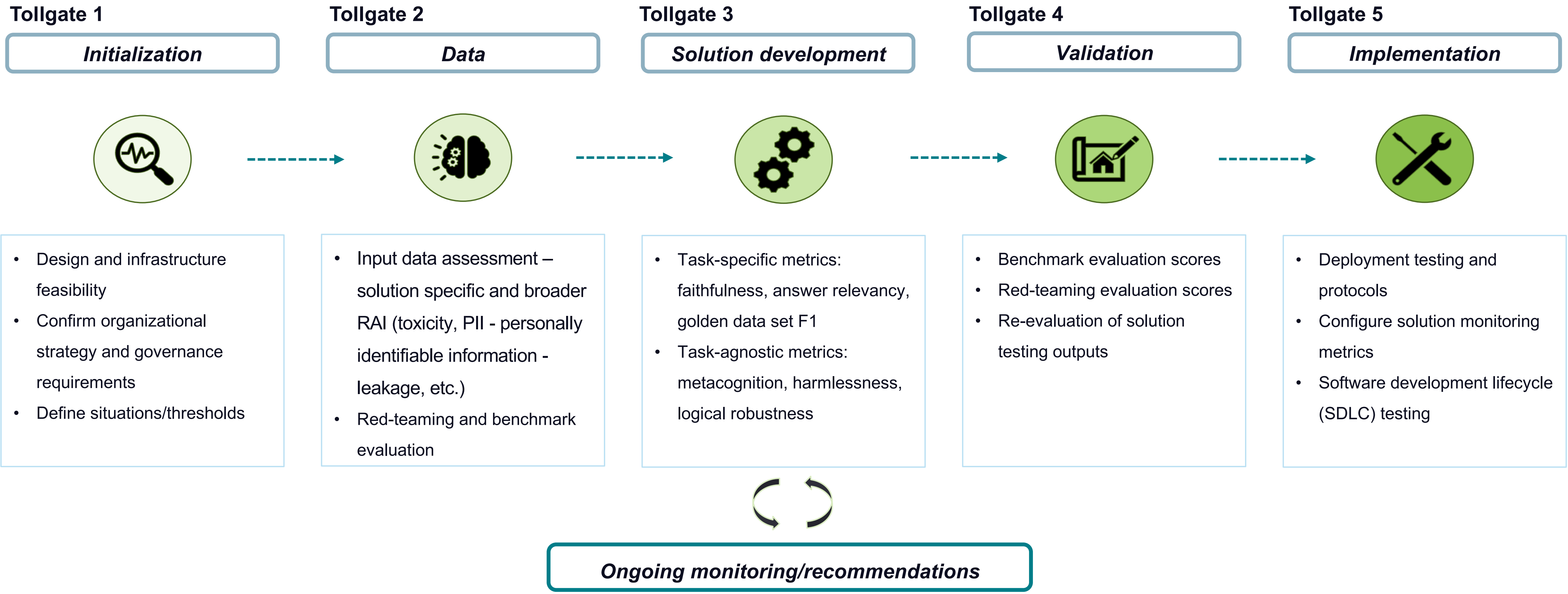
Key AI control principles establish the foundational principles for AI agnostic to the underlying use case/technique.



Key Control Principles

1. Validation
2. Verification
3. Preventive controls
4. Operational resiliency/ stress testing
5. Ecosystem monitoring
6. Human control/override
7. Rules, regulations and laws
8. Organization’s code of conduct

Need for Tollgates at the Onset and Throughout the Generative AI Development Lifecycle



Tollgate 1

Initialization

- Design and infrastructure feasibility
- Confirm organizational strategy and governance requirements
- Define situations/thresholds

Tollgate 2

Data

- Input data assessment – solution specific and broader RAI (toxicity, PII - personally identifiable information - leakage, etc.)
- Red-teaming and benchmark evaluation

Tollgate 3

Solution development

- Task-specific metrics: faithfulness, answer relevancy, golden data set F1
- Task-agnostic metrics: metacognition, harmlessness, logical robustness

Tollgate 4

Validation

- Benchmark evaluation scores
- Red-teaming evaluation scores
- Re-evaluation of solution testing outputs

Tollgate 5

Implementation

- Deployment testing and protocols
- Configure solution monitoring metrics
- Software development lifecycle (SDLC) testing

Ongoing monitoring/recommendations

Evaluation Considerations for Generative AI Solutions to Ensure Performance, and Identify and Mitigate Risks

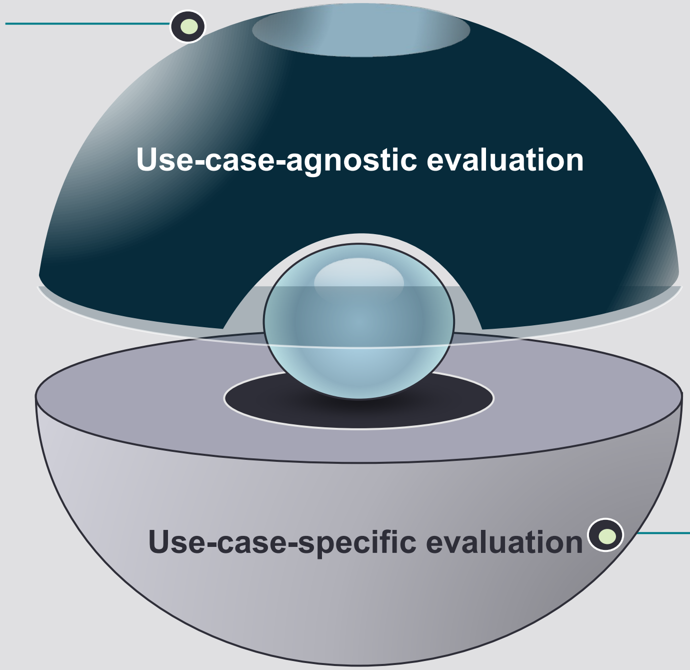
OVERALL SOLUTION EVALUATION

LLM SOLUTION-LEVEL EVALUATION

Use-case-agnostic evaluation

Assess pretrained large language models (LLMs) for suitability, performance against specific responsible AI consideration, and vulnerabilities through benchmarking, red-teaming, etc.

- Evaluation against benchmark datasets
- Vulnerability testing for adverse scenarios
- Comparative assessment across alternatives



Use-case-specific evaluation

Specific evaluation based on the business solution and task, intended to establish trackable metrics within each stage of the solution lifecycle and assess risk exposure

- Input evaluation
- Solution design and performance evaluation
- Output evaluation and ongoing monitoring

Data quality testing

Unit testing

Functional testing

Security testing

Integration testing

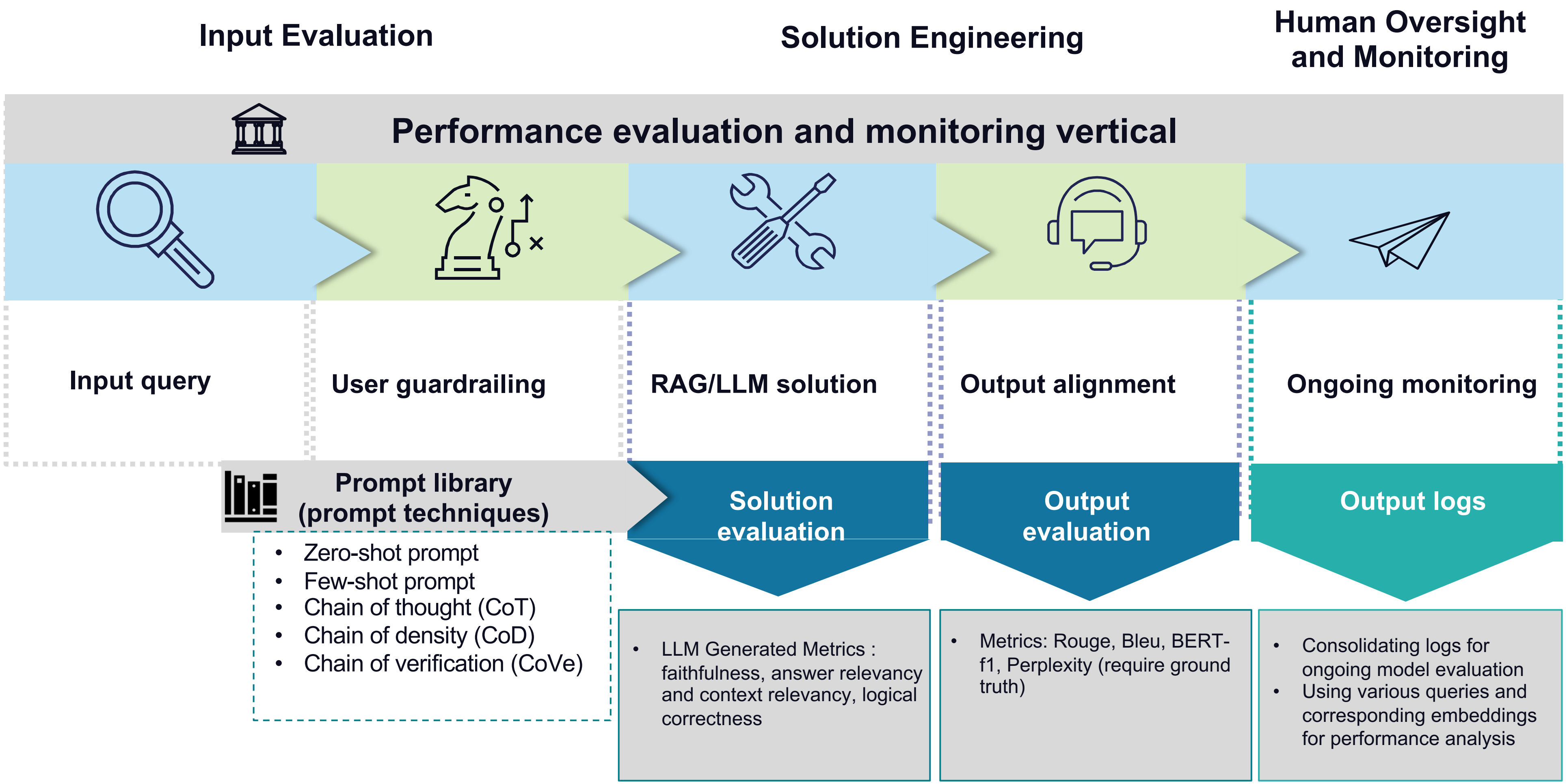
User acceptance testing

Usability testing

Continuous monitoring

Maintenance planning

Use-case Specific Solution Design and Evaluation Consideration



Input Evaluation

Solution Engineering

Human Oversight and Monitoring

Performance evaluation and monitoring vertical

Input query

User guardrailing

RAG/LLM solution

Output alignment

Ongoing monitoring

Prompt library (prompt techniques)

- Zero-shot prompt
- Few-shot prompt
- Chain of thought (CoT)
- Chain of density (CoD)
- Chain of verification (CoVe)

Solution evaluation

- LLM Generated Metrics : faithfulness, answer relevancy and context relevancy, logical correctness

Output evaluation

- Metrics: Rouge, Bleu, BERT-f1, Perplexity (require ground truth)

Output logs

- Consolidating logs for ongoing model evaluation
- Using various queries and corresponding embeddings for performance analysis

Select Prompting Techniques to Build Robust LLM Solutions

Zero-shot prompting

Prompt

Write a summary of the following news article. Article: X

Output:
Summary: …

Few-shot prompting

Prompt

Write a summary of the following news article.
Article: X
Here is a sample:
Article: "A study says eating chocolate weekly lowers heart disease ... eaten in moderation."
Summary: The study links chocolate consumption … best choice, but moderation is key

Output:
Summary: …

Chain-of-thought prompting

Prompt

Write a summary of the following news article.
Article: X
Instructions: Outline the thought process step by step

Output:
Step 1: Fact X is essential …
Step 2: Following reasoning depends on fact …
Summary: …

Chain-of-density prompting

Prompt

Generate concise, entity-dense summaries of the Article: X
Identify up to three informative entities missing from the previous summary. Write a new, denser summary covering all entities and details from the previous summary plus the missing entities.

Output:
Summary 1: …
Summary 2: …

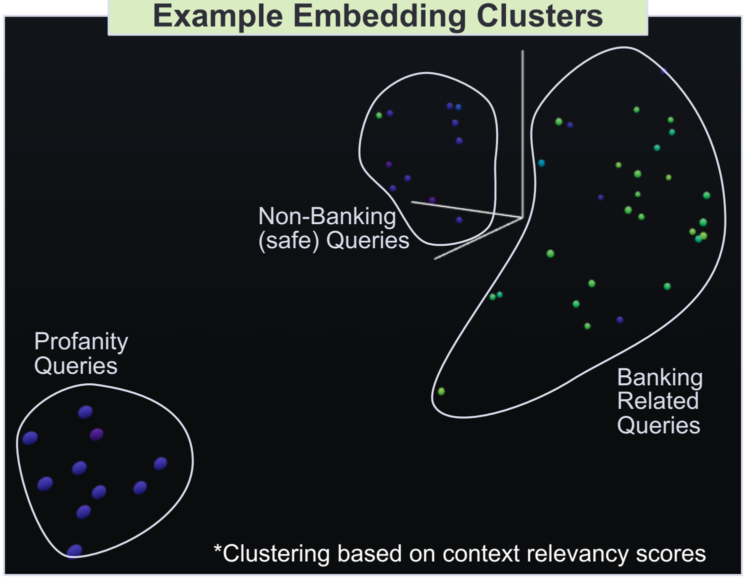
Ongoing Monitoring of Developed Solution to Ensure Performance

Accessible visualizations and statistics can be employed for ongoing monitoring & human-oversight. These methods can help adding efficiencies to the manual review requirements for tracking the solution post launch.

Approach

- ❑ Create vector embeddings of the knowledge base, query, and response
- ❑ Reduce dimensionality of the vector representation using UMAP* for enhanced visualization
- ❑ Perform data clustering using HDBSCAN** to automatically cluster data points
- ❑ Introduce additional metrics to enhance visualization capabilities
- ❑ Easily surface up and filter on clusters which exhibit unique characteristics for ongoing monitoring and human-oversight

Example Embedding Clusters



*Clustering based on context relevancy scores

*UMAP: Uniform Manifold Approximation and Projection

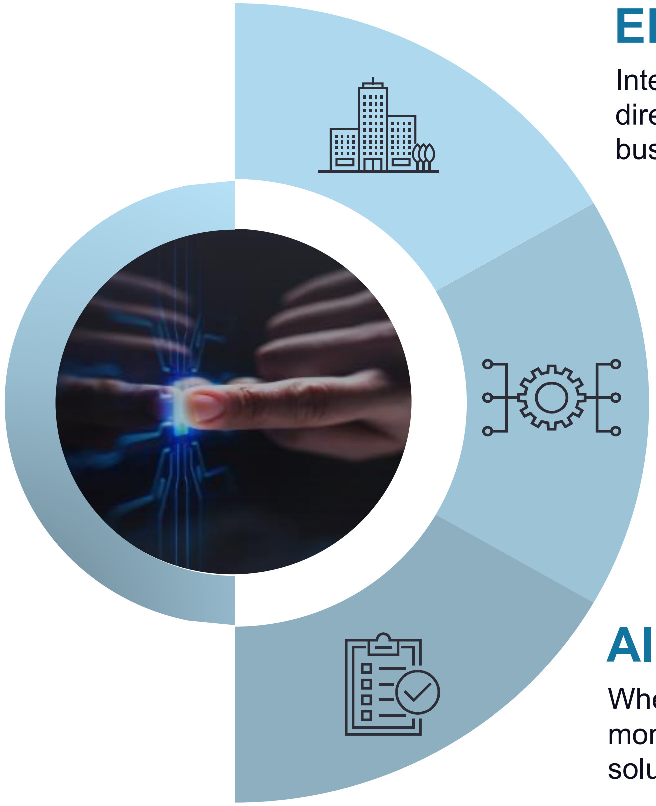
**HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise

Driving Value Through AI

The evolving regulatory and compliance landscape necessitates robust AI governance at enterprise, process and use case levels to mitigate risks, enforce controls and sustain value creation through AI.

AI VALUE CREATION

Responsible AI governance across levels drives economic, social and organizational benefits by enabling ethical compliance, fostering innovation, and supporting sustainable, long-term value creation in line with societal expectations.



ENTERPRISE LEVEL

Integrate an AI governance framework at the enterprise level to set strategic direction and policies for AI utilization, facilitating ethical practices and alignment with business strategy.

PROCESS LEVEL

At the process level, it is essential to identify risks and implement controls to maintain the integrity and governance of AI operations.

AI USE CASE LEVEL

When developing AI use cases, it’s crucial to design safe AI solutions with built-in monitoring protocols and to incorporate independent validation checks to maintain solution integrity and enable responsible use of AI.