Coalition for Health AI (CHAI): Ensuring Trustworthy AI in Health

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Introduction

Who We Are

- **Over 800+ Private Sector Organizations:** Health Systems, Payors, Device Manufacturers, Technology Companies, Patient Advocates
- **Founding Members:** Mayo Clinic, Duke Health, MITRE, UC Berkeley, Johns Hopkins, Stanford Medicine, UCSF & UC Health
- **Industry Partners:** Optum, Google, Microsoft, SAS
- **US Govt Partners:** HHS, FDA, ONC, NIH, CMS, White House OSTP, AHRQ, VA, NIST, CDC

Vision & Mission Statement

Our Vision is to have more **trustworthy** and **transparently** developed and maintained Health AI.

Our Mission is to provide a **framework** for the landscape of health AI tools to ensure high quality care, increase trust amongst users, and meet health care needs.

BLUEPRINT FOR TRUSTWORTHY AI IMPLEMENTATION GUIDANCE AND ASSURANCE FOR HEALTHCARE

COALITION FOR HEALTH AI

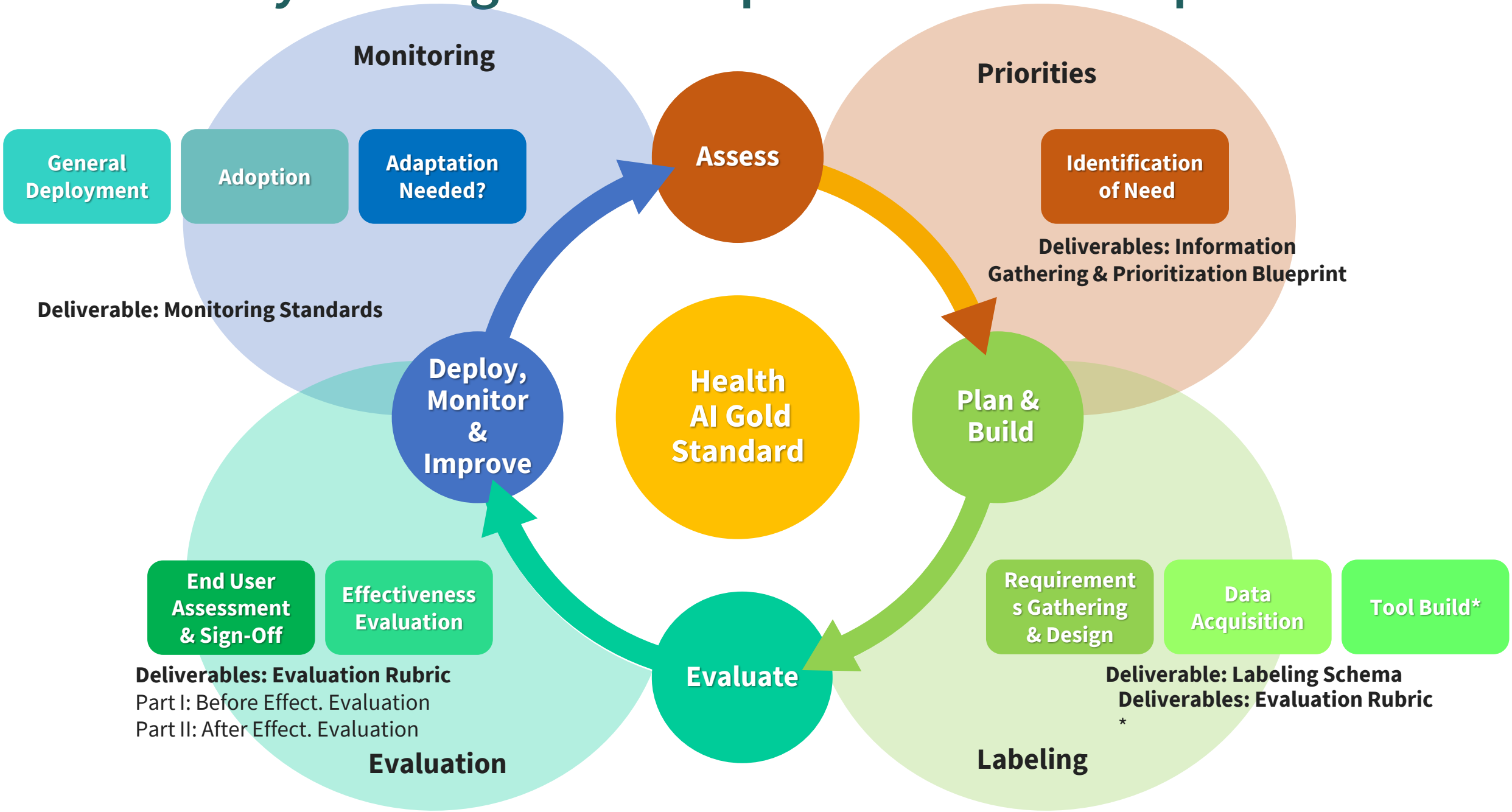
VERSION 1.0 _ APRIL 04, 2023

Core Principles

Aligned to NIST AI Risk Management Framework and the White House Blueprint for an AI Bill of Rights

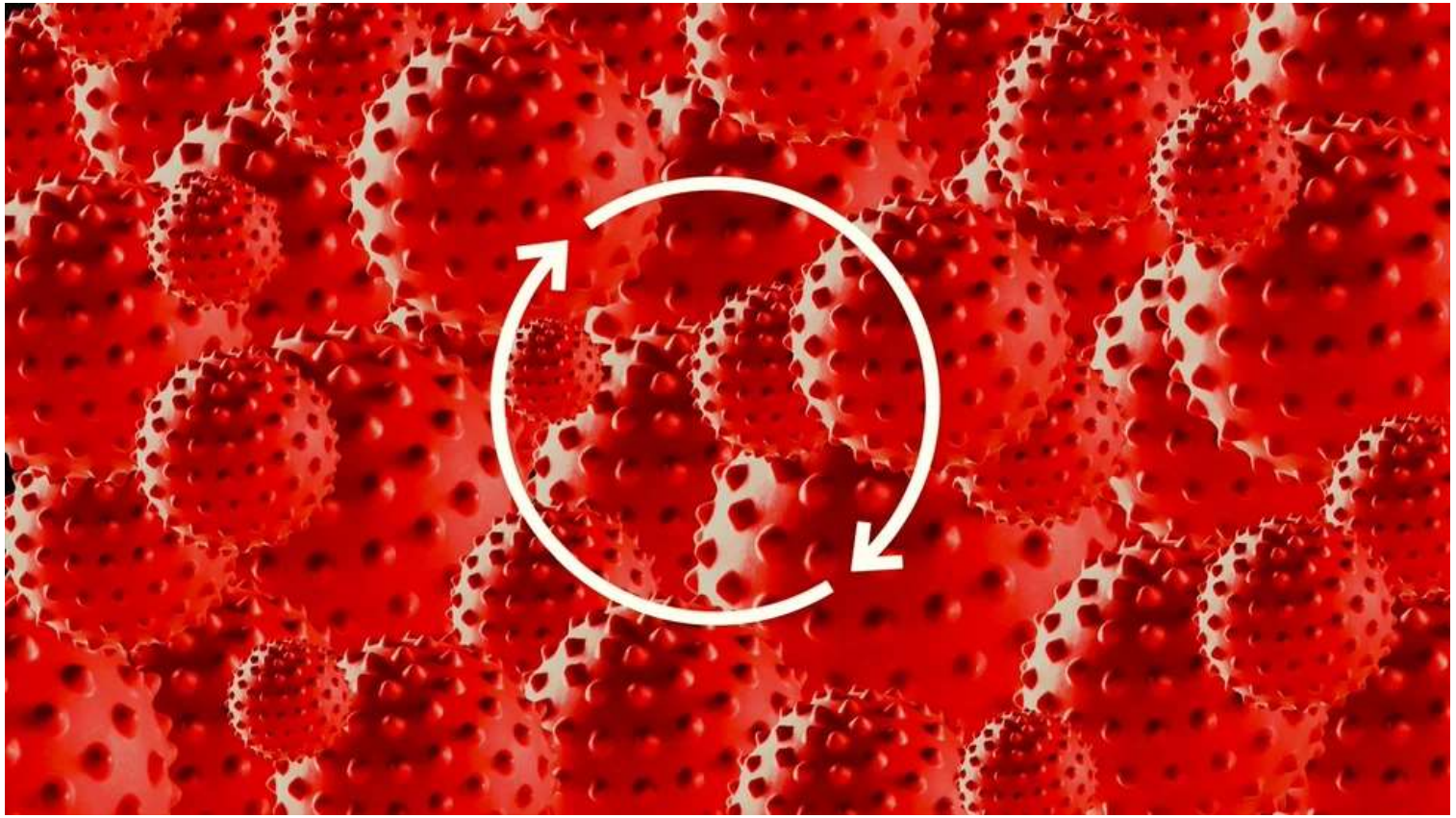
MITRE

Lifecycle Stages & ML Ops Toolkit Development



“Technological solutions tend to rise into society’s penthouses, while epidemics seep into its cracks.”

- Ed Yong, *The Atlantic*



How AI Is Changing Healthcare

All Models are Local

Health AI Assurance Labs

GPT-4 System Card

OpenAI

March 23, 2023

Abstract

Large language models (LLMs) are being deployed in many domains of our lives ranging from browsing, to voice assistants, to coding assistance tools, and have potential for vast societal impacts.[1, 2, 3, 4, 5, 6, 7] This system card analyzes GPT-4, the latest LLM in the GPT family of models.[8, 9, 10] First, we highlight safety challenges presented by the model's limitations (e.g., producing convincing text that is subtly false) and capabilities (e.g., increased adeptness at providing illicit advice, performance in dual-use capabilities, and risky emergent behaviors). Second, we give a high-level overview of the safety processes OpenAI adopted to prepare GPT-4 for deployment. This spans our work across measurements, model-level changes, product- and system-level interventions (such as monitoring and policies), and external expert engagement. Finally, we demonstrate that while our mitigations and processes alter GPT-4's behavior and prevent certain kinds of misuses, they are limited and remain brittle in some cases. This points to the need for anticipatory planning and governance.[11]

Content Warning: This document contains content that some may find disturbing or offensive, including content that is sexual, hateful, or violent in nature.

An Urgent Need to Rethink How We Regulate LLMs