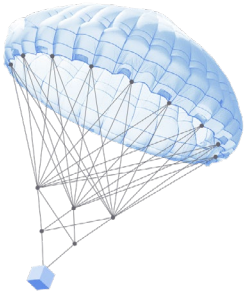
PARTNERSHIP ON AI



# PAI's Guidance for Safe Foundation Model Deployment

## A Framework for Collective Action

Partnership on AI's Guidance for Safe Foundation Model Deployment is a framework for model providers to responsibly develop and deploy a range of AI models, promote safety for society, and adapt to evolving capabilities and uses.

**We are actively seeking public comments on the Model Deployment Guidance until January 15, 2024, and plan to release an updated version later in the year. Please use the Model Deployment Guidance website to submit your feedback.**

## About the Guidance

Recent years have seen rapid advances in AI driven by foundation models, sometimes known as large language models or general purpose AI. These are transformative AI systems trained on large datasets which power a variety of applications, from content generation to interactive conversational interfaces. Already, there is widespread recognition of this technology's potential for both social benefit and harm. The use of foundation models could enable new forms of creative expression, boost productivity, and accelerate scientific discovery. It could also increase misinformation, negatively impact workers, and automate criminal activity.

Already, there is widespread recognition of this technology's potential for both social benefit and harm.

Given the potentially far-reaching impacts of foundation models, shared safety principles must be translated into practical guidance for model providers. This requires collective action. To establish effective, collectively-agreed upon practices for responsible model development and deployment, diverse voices across industry, civil society, academia, and government need to work together.

In collaboration with our global community of civil society, industry, and academic organizations, **PAI is releasing our Guidance for Safe Foundation Model Deployment for public comment on our website.** This is a framework for model providers to responsibly develop and deploy foundation models across a spectrum of current and emerging capabilities, helping anticipate and address risks. The Model Deployment Guidance gives AI developers practical recommendations for operationalizing AI safety principles. Through the public comment process, even more stakeholders will help shape this truly collective effort.

PRACTICAL GUIDANCE FOR FOUNDATION MODEL SAFETY

CREATED THROUGH ONGOING MULTISTAKEHOLDER COLLABORATION

CUSTOMIZABLE FOR SPECIFIC MODEL AND RELEASE TYPES

DESIGNED TO EVOLVE AS NEW CAPABILITIES AND RISKS EMERGE

partnershiponai.org/modeldeployment

Using PAI's Model Deployment Guidance website, foundation model providers can receive a set of recommended practices to follow throughout the deployment process, tailored to the capabilities of their specific model and how it is being released. **Designed to be a living document that can appropriately respond to new capabilities as AI technologies continue to evolve,** PAI's Model Deployment Guidance aims to complement broader regulatory approaches.

**Foundation model providers can receive a set of recommended practices to follow throughout the deployment process.**

## Key Features of the Guidance

The Model Deployment Guidance's guidelines establish a normative baseline and suggest additional practices for responsible development of foundation models, allowing collaborative reassessment as capabilities and uses advance. This accommodates diverse AI models and deployment scenarios. Not intended as a comprehensive set of instructions for implementation, these guidelines provide a framework for ongoing collective research and action. The guidelines aim to inform and catalyze other individual and collaborative efforts to develop specific guidance or tooling in alignment with the guidelines.

- **Scaling oversight and safety**
  To address risks appropriately, the Model Deployment Guidance's guidelines are tailored to scale oversight and safety practices based on the capabilities and availability of each AI model. The Model Deployment Guidance avoids oversimplification by not solely equating model size or generality with risk.
- **Open access guidance**
  The Model Deployment Guidance includes guidelines for open access models, offering a starting point into transparency and risk mitigation strategies. This provides guidance for both current and future providers of open source models.
- **Broad applicability**
  The Model Deployment Guidance applies across the spectrum of foundation models, from existing to frontier.
- **Cautious frontier model rollout**
  The Model Deployment Guidance recommends staged releases and restricted access for frontier models initially until adequate safeguards are demonstrated.
- **Holistic view of safety**
  The Model Deployment Guidance establishes starting points to address a wide variety of safety risks, including potential harms related to bias, overreliance on AI systems, worker treatment, and malicious activities by bad actors.

# The Guidelines

Below is a list of a list of the Model Deployment Guidance's 22 possible guidelines. Please visit the Model Deployment Guidance website to generate a tailored set of guidelines for different model and release types, including a complete list of baseline and recommended practices for each guideline.

## Research & Development

| | | |
|---|---|---|
| 1 | **Scan for novel or emerging risks** | Proactively identify and address potential novel or emerging risks from foundation/ frontier models. |
| 2 | **Practice responsible iteration** | Practice responsible iteration to mitigate potential risks when developing and deploying foundation/frontier models, through both internal testing and limited external releases. |
| 3 | **Assess upstream security vulnerabilities** | Identify and address potential security vulnerabilities in foundation/frontier models to prevent unauthorized access or leaks. |
| 4 | **Produce a "Pre-Systems Card"** | Disclose planned testing, evaluation, and risk management procedures for foundation/frontier models prior to development. |
| 5 | **Establish risk management and responsible AI structures for foundation models.** | Establish risk management oversight processes and continuously adapt to address real world impacts from foundation/frontier models. |

## Pre-Deployment

| | | |
|---|---|---|
| 6 | **Internally evaluate models for safety** | Perform internal evaluations of models prior to release to assess and mitigate for potential societal risks, malicious uses, and other identified risks. |
| 7 | **Conduct external model evaluations to assess safety** | Complement internal testing through model access to third-party researchers to assess and mitigate potential societal risks, malicious uses, and other identified risks. |
| 8 | **Undertake red-teaming and share findings** | Implement red teaming that probes foundation/frontier models for potential malicious uses, societal risks and other identified risks prior to release. Address risks and responsibly disclose findings to advance collective knowledge. |
| 9 | **Publicly report model impacts and "key ingredient list"** | Provide public transparency into foundation/frontier models' "key ingredients" testing evaluations, limitations and potential risks to enable cross-stakeholder exploration of societal risks and malicious uses. |
| 10 | **Provide downstream use documentation** | Equip downstream developers with comprehensive documentation and guidance needed to build safe, ethical, and responsible applications using foundation/frontier models.<br>*(Note: It is well understood downstream developers play a crucial role in anticipating deployment-specific risks and unintended consequences. This guidance aims to support developers in fulfilling that responsibility.)* |
| 11 | **Establish safeguards to restrict unsafe uses** | Implement necessary organizational, procedural and technical safeguards, guidelines and controls to restrict unsafe uses and mitigate risks from foundation/frontier models. |

## Post-Deployment

| | | |
|---|---|---|
| **12** | **Monitor deployed systems** | Continuously monitor foundation/frontier models post-deployment to identify and address issues, misuse, and societal impacts. |
| **13** | **Implement incident reporting** | Enable timely and responsible reporting of safety incidents to improve collective learning. |
| **14** | **Establish decommissioning policies** | Responsibly retire foundation/frontier models from active use based on well-defined criteria and processes. |
| **15** | **Develop transparency reporting standards** | Collaboratively establish clear transparency reporting standards for disclosing foundation/frontier model usage and policy violations. |

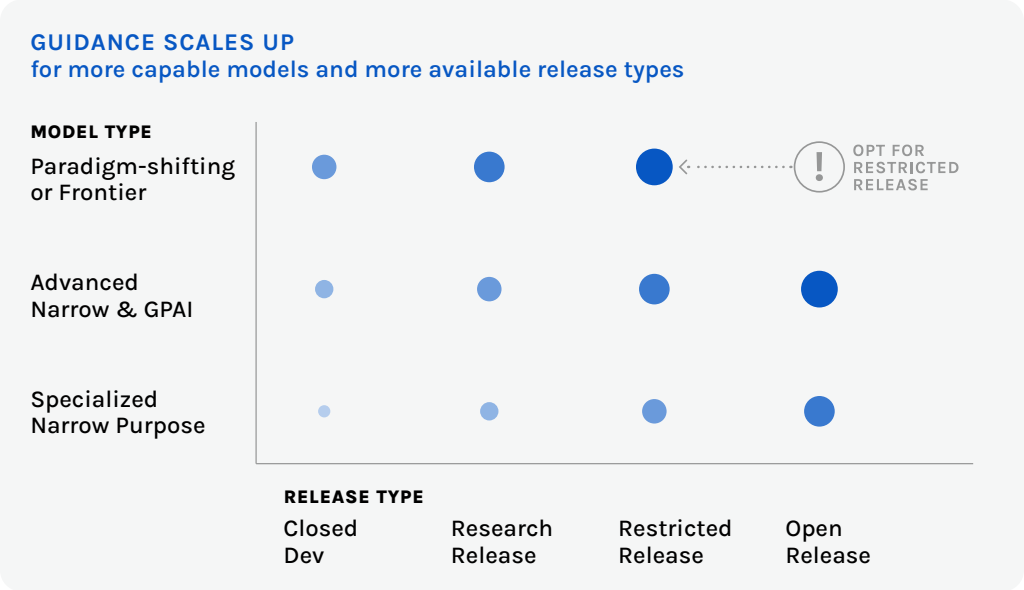
## Societal Impact

**(cross-cutting through the model's lifecycle)**

| | | |
|---|---|---|
| **16** | **Support third party inspection of models and training data** | Support progress of third-party auditing capabilities for responsible foundation/frontier model development through collaboration, innovation and transparency. |
| **17** | **Responsibly source all labor including data enrichment** | Responsibly source all forms of labor, including for data enrichment tasks like data annotation and human verification of model outputs. |
| **18** | **Conduct human rights due diligence** | Implement comprehensive human rights due diligence methodologies to assess and address the impacts of foundation/frontier models. |
| **19** | **Enable feedback mechanisms across the AI value chain** | Implement inclusive feedback loops across the AI value chain to ethically identify potential harms. |
| **20** | **Measure and disclose environmental impacts** | Measure and disclose the environmental impacts resulting from developing and deploying foundation/frontier models. |
| **21** | **Disclose synthetic content** | Adopt responsible practices for disclosing synthetic media and advance solutions for identifying other synthetic content |
| **22** | **Measure and disclose anticipated severe labor market risks** | Measure and disclose potential severe labor market risks from deployment of foundation/frontier models. |

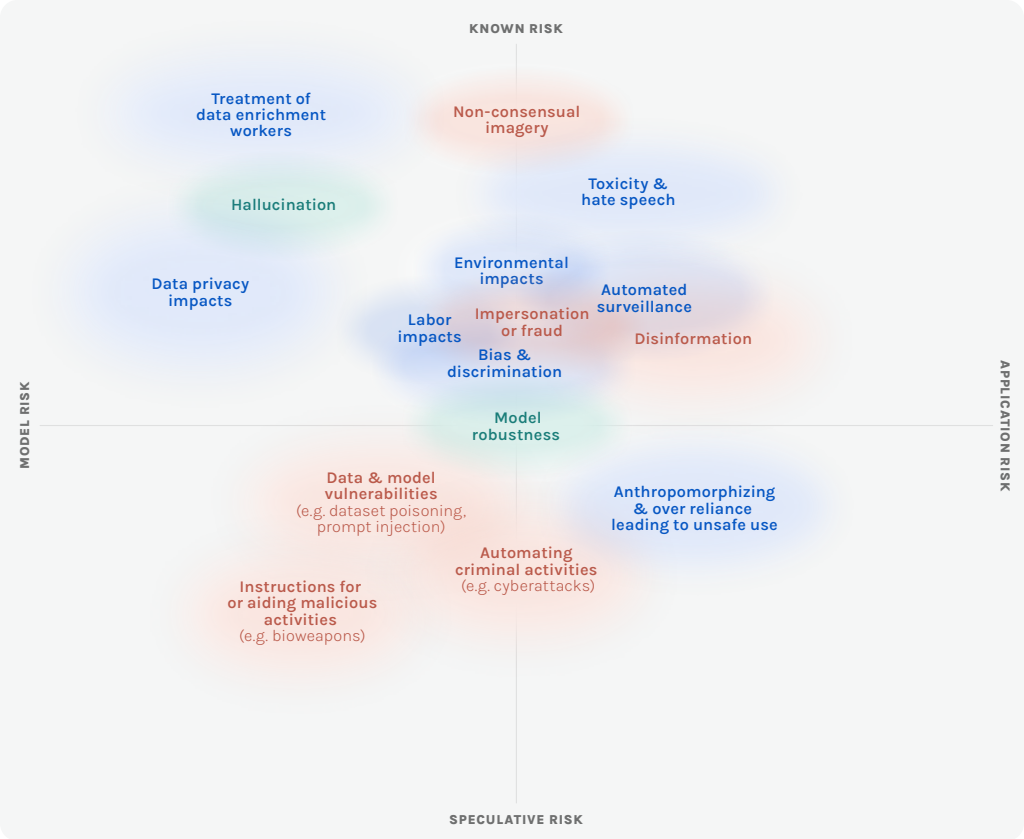
## How the Guidance Tailors to Different Model and Release Types

There are a total of 22 possible guidelines included in the Model Deployment Guidance. Not all model and release types are treated the same within the paradigm of the Model Deployment Guidance. **The suggested guidelines are more extensive for more capable models and more available release types.** The full 22 guidelines apply to the "Frontier and Restricted" model and release category. This concept is visualized below:



# Safety Risks Addressed by the Guidance

The Model Deployment Guidance addresses both risks from the foundation models themselves and risks that can arise downstream when others build applications using the models. While downstream developers have an important role in managing application risks, in this guidance, model providers adopt accountability measures like providing synthetic media disclosures and supplying downstream use documentation, thereby addressing select application risks within the scope of the guidance.




SUB-CATEGORIES OF RISKS

- **Malicious uses:** Risks of intentional misuse or weaponization of models to cause harm
- **Societal risks:** Potential harms that negatively impact society, communities and groups
- **Other Risks:** Risks distinct from the above categories

**Model risk** refers to the potential risks associated with the foundation model itself. Includes biases in the training data, human-computer interaction harms resulting from interacting with the model, or vulnerabilities to adversarial attacks. Model risks focus on the inherent characteristics of the model and other negative impacts that model providers can address.

**Application risk** refers to potential risks that arise from downstream use-cases and applications built using foundation models or when these models are integrated into real-world products and services. Includes potential harms caused by incorrect or biased outputs and malicious uses.

**Known risks** are the risks that have been identified, acknowledged, and are reasonably well-understood. These risks are typically based on empirical evidence, research, or previous experiences with similar models or applications. Known risks are usually more predictable and quantifiable.

**Speculative risks** are the risks that are uncertain, hypothetical, or potential but have not been observed repeatedly or thoroughly studied. These risks may arise from emerging technologies, complex interactions, or unexpected consequences that are difficult to anticipate. Speculative risks are often more challenging to quantify or mitigate due to their uncertain nature.

This illustration doesn't cover all risks associated with foundation models or their secondary and tertiary societal impacts. The risk landscape will evolve over time based on input from stakeholders and technological advances.

# Definitions

### FOUNDATION MODEL

The Model Deployment Guidance uses "foundation model" to encompass all models with generally applicable functions that are designed to be used across a variety of contexts. The current generation of these systems is characterized by training deep learning models on large datasets (which requires significant computational resources) to perform numerous tasks that can serve as the "foundation" for a wide array of downstream applications.

### MODEL PROVIDERS

The Model Deployment Guidance distinguishes model providers from actors in the broader AI ecosystem (seen below) as those training foundational models that others may build on.

| ECOSYSTEM ACTOR | ROLE DESCRIPTION |
|---|---|
| Compute / Hardware Providers | Providing underlying compute power to train and run models |
| Cloud Providers | Providing underlying cloud infrastructure to support training of and deployed models |
| Data Providers | Providing training datasets (intentionally or unintentionally) for model providers, may also be model providers |
| **Model Providers** | **Training foundational models (proprietary or open-source) that others may build on** |
| Application Developers (or: Service Developers, Model Integrators) | Building applications and services on top of foundational models |
| Consumers and/or Affected Users | Consumers (B2C) who are end-users of services built on top of foundational models<br>Affected Users may be impacted or implicated in the use of AI (e.g. medical AI Consumers are doctors, and Affected Users are patients) |

# What's Next for the Guidance

As PAI continues development of the Model Deployment Guidance, we seek public comment to incorporate into an updated version we plan to release in 2024. Our next steps include the following key actions.

- **Collaborative Group**
  We will bring together a collaborative group focused on applying the Framework in practice through yearly case examples or analysis via a public reporting process. This will help us identify challenges and trade-offs that may arise, and we'll share our findings.
- **Operationalization Support**
  We will provide tactical options to put our key guidelines into operation. We aim to support the implementation of these guidelines over time to ensure they are effective.
- **Shared Responsibility**
  We will explore how responsibility should be shared across the evolving value chain for foundation models.
- **Regular Updates**
  We will continue to update our model and release categorization, ensuring that it remains current and relevant to the evolving landscape.