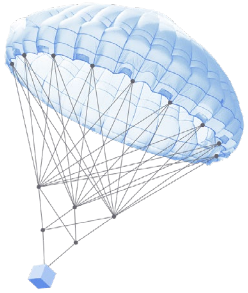
PARTNERSHIP ON AI



PAI's Guidance for Safe Foundation Model Deployment

MODEL TYPE

Paradigm-shifting or Frontier

RELEASE TYPE

Closed Development

Partnership on AI's Guidance for Safe Foundation Model Deployment, first released in 2023, provides a framework for model providers to responsibly develop and deploy AI models. Following extensive stakeholder feedback during our public comment period highlighting the need to examine roles and responsibilities across the AI value chain, PAI has produced expanded guidance addressing key actors beyond model providers, including model adapters, model hosting services, and application developers, with particular focus on open foundation models.

Recent advances in foundation models have transformed the AI landscape, enabling content generation and paving the way for interactive systems that will be capable of performing complex digital tasks autonomously. While these models offer unprecedented opportunities for scientific discovery, productivity enhancement, and creative expression, they also present complex challenges including potential misuse, novel risks from increasingly capable systems, and the need for robust safety measures.

The Model Deployment Guidance website provides guidelines for various model capability and release type combinations. The Guidance also addresses significant model updates that expand capabilities post-deployment, requiring renewed governance processes.

The guidelines scale according to model capabilities and release types, with more extensive requirements for more capable models and widely available releases. This framework is meant to inform emerging regulatory frameworks, including the EU's general purpose AI Code of Practice, while providing practical guidance for safety measures companies should invest in developing.

Given the potentially far-reaching impacts of foundation models, translating shared safety principles into practical guidance requires collective action. These frameworks represent ongoing collaboration between industry, civil society, academia, and government to establish effective, collectively-agreed upon practices for responsible AI development and deployment.

This Guidance Checklist is one of three targeted frameworks that address distinct development and deployment scenarios.

Please see the other two frameworks:

Frontier x Restricted Release

For paradigm-shifting foundation models requiring extensive safety measures

General Purpose x Open Release

Decentralized approaches emphasizing collaborative value chain governance

Definitions

FOUNDATION MODELS

Foundation models are large-scale base models trained on vast amounts of data, capable of being adapted to a wide range of downstream tasks through methods like fine-tuning or prompting. These models, also known as "general purpose AI," serve as starting points for developing more specialized AI systems across scientific and commercial domains. Increasingly, these models are being integrated into operating systems and services as AI assistants or "agents," capable of understanding personal context and eventually performing complex tasks across applications.

MODEL PROVIDERS

Model providers are organizations that train foundation models and distribute their components (such as model weights), that others may build on. These providers may operate with different objectives and distribution approaches:

- Research purposes (enabling scientific investigation and advancement)
- Open source development (allowing free access, modification, and distribution)
- Commercial provision (offering paid services or products)

Guidance Checklist

MODEL TYPE

Paradigm-shifting or Frontier

Cutting edge general purpose models that significantly advance capabilities across modalities compared to the current state of the art.

KEY CONSIDERATIONS

A model could be considered frontier or paradigm-shifting if it meets at least some of these criteria:

- Enables capabilities that significantly advance the current state-of-the-art
- Demonstrates breakthrough scaling in parameters or computational resources beyond current standards
- Shows self-learning capabilities exceeding current AI systems
- Enables direct real-world actions beyond passive information processing, or 'agentic AI', through released interfaces or applications

RELEASE TYPE

Restricted API and Hosted Access

Models available only through a controlled API, cloud platform, or hosted through a proprietary interface, with limits on use. Does not provide direct possession of the model. Allows restricting access and monitoring usage to reduce potential harms.

Research & Development

Scan for novel or emerging risks	Proactively identify and address potential novel or emerging risks from foundation/frontier models.
BASELINE PRACTICES	Conduct model evaluations and experiments to identify new indicators for novel risks, including potential negative societal impacts, malicious uses, "dangerous capabilities" like persuasion and other speculative risks. Study potential risks from integrating the model into novel or unexpected downstream environments and use cases. Assess their likelihood and potential impact. Establish regular processes to probe and address potential novel or emerging risks through techniques like external red teaming.
RECOMMENDED PRACTICES	Collaborate with relevant stakeholders (ie. governments, other labs and academic researchers) to advance the identification of novel risks and responsible disclosure practices.
IMPLEMENTATION RESOURCES	Organizations such as the Frontier Model Forum can contribute to the ongoing development of novel risk assessment practices.
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. *(This is an emerging and less-explored guardrail, and disclosure practices will need to evolve as we learn more about its effectiveness.)*
BASELINE PRACTICES	Before model development, forecast intended capabilities and likely outcomes to inform risk assessments. Commence model development on a smaller scale, systematically test for risks during internal iterations through evaluations and red teaming, incrementally address identified risks, and update forecasts of model capabilities and risks (internal iteration). Deploy frontier models in limited, experimental environments to study impacts before considering full deployment (external iteration). Adapt deployment based on risk assessments and learnings during iterations.

RECOMMENDED PRACTICES	Maintain documentation of each iteration, including lessons learned, challenges faced, and mitigation measures implemented. Share documentation with government bodies as required and seek feedback from a diverse range of stakeholders including domain experts and affected users to inform the iteration process and risk mitigation strategies. Collaborate with stakeholders to inform and advance responsible iteration approaches.
Assess upstream security vulnerabilities	Identify and address potential security vulnerabilities in foundation/frontier models to prevent unauthorized access or leaks.
BASELINE PRACTICES	Implement comprehensive cybersecurity standards at the start of the development. Conduct rigorous testing such as penetration testing, prompt analysis, and data poisoning assessments to identify vulnerabilities that could enable model leaks or manipulation. Establish protocols for addressing identified vulnerabilities pre-deployment.
RECOMMENDED PRACTICES	Exceed baseline cybersecurity standards as risks and use cases evolve, drawing on guidance from standards bodies. Offer bug bounty programs to encourage external vulnerability discovery. Share lessons learned across industry to collectively strengthen defenses. Release regular updates to the model that patches security vulnerabilities.
IMPLEMENTATION RESOURCES	UC Berkeley CLTC Draft Profile under Measure 2.7
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.
BASELINE PRACTICES	Establish risk management structures and processes,such as enterprise risk management, independent safety boards, and ethics review processes to define guidelines on responsible development, release, and staged rollout considerations, including when a model should not be released. Regularly update policies, frameworks, and organizational oversight to address evolving capabilities and real-world impacts.
RECOMMENDED PRACTICES	Publicly share information about implemented internal governance and risk management processes.

Pre-Deployment

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.
BASELINE PRACTICES	Establish comprehensive internal evaluation policies and processes including testing for fairness, interpretability, output harms, novel risks and intended vs foreseeable unintended use cases. Proactively identify and minimize potential sources of bias in training corpora, and adopt techniques to minimize unsafe model behavior. Conduct evaluations using cross-disciplinary review teams spanning ethics, security, social science, safety and other relevant domains. Address identified risks and adapt deployment plans accordingly based on learnings from pre-deployment evaluations. Maintain documentation of evaluation methods, results, limitations, and steps taken to address identified issues and integrate insights in public reporting per guidance below.
RECOMMENDED PRACTICES	Collaborate across industry, civil society, and academia to advance the development and standardization of model evaluations for foundation models.
IMPLEMENTATION RESOURCES	Organizations such as Frontier Model Forum can contribute to the ongoing development of internal model evaluations.

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.
BASELINE PRACTICES	Perform internal and external red teaming across model capabilities, use cases, and potential harms including dual-use risks using techniques such as adversarial testing, vulnerability scanning, and surfacing edge cases and failure modes. Conduct iterative red teaming throughout model development. Continuously evaluate results to identify areas for risk mitigation and improvements, including for planned safeguards. Commission external red teaming by independent experts such as domain experts and affected users to surface gaps. Select external red teamers to incentivize the objective discovery of flaws and ensure adequate independence. Address identified risks and adapt deployment plans accordingly based on learnings from pre-deployment evaluations. Responsibly disclose findings, aligned with guidance below on public reporting.
RECOMMENDED PRACTICES	Collaborate across industry, civil society, and academia to advance red teaming methodologies and responsible disclosures.

Societal Impact (cross-cutting through the model's lifecycle)

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.
BASELINE PRACTICES	Pay or contract with vendors that will pay data enrichment workers above the workers' local living wage. Provide or contract with vendors that provide clear instructions for enrichment tasks that are tested for clarity. Enable workers to opt out of tasks. Equip or contract with vendors that equip workers with simple and effective mechanisms for reporting issues, asking questions, and providing feedback on the instructions or task design.
RECOMMENDED PRACTICES	Design and run a pilot before launching a data enrichment project. Disclose any new types of labor that enter the supply chain of foundation models. Ensure policies and responsible sourcing practices extend as appropriate to new labor sources as they emerge, like red teamers. Update internal standards and vendor agreements accordingly. Proactively survey all workers to identify areas for improving policies, instructions, and work environments, and seek external feedback.
IMPLEMENTATION RESOURCES	PAI's Library of Practitioner Resources for responsible data enrichment sourcing.
Measure and disclose environmental impacts	Measure and disclose the environmental impacts resulting from developing and deploying foundation/frontier models.
BASELINE PRACTICES	Establish processes to evaluate environmental costs like energy usage, carbon emissions and other metrics. Monitor and report on environmental impacts of model development and deployment.
RECOMMENDED PRACTICES	Provide environmental measurement/disclosure mechanisms for application developers building on frontier models. Incorporate impacts into model development decisions. Collaborate across industry, civil society, and academia to advance the measurement of environmental impacts and responsible disclosure practices.