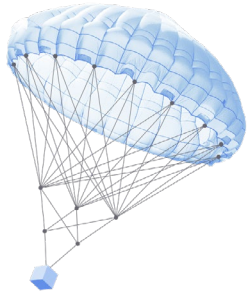
PARTNERSHIP ON AI



# PAI's Guidance for Safe Foundation Model Deployment

**MODEL TYPE**
Paradigm-shifting or Frontier

**RELEASE TYPE**
Restricted API and Hosted Access

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:

General Purpose x Open Release

Decentralized approaches emphasizing collaborative value chain governance

Frontier x Closed Deployment

For internal deployments where models are directly integrated into products without public release

## 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.<br><br>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.<br><br>*(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.<br><br>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).<br><br>Deploy frontier models in limited, experimental environments to study impacts before considering full deployment (external iteration).<br><br>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.<br><br>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.<br><br>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.<br><br>Conduct rigorous testing such as penetration testing, prompt analysis, and data poisoning assessments to identify vulnerabilities that could enable model leaks or manipulation.<br><br>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.<br><br>Offer bug bounty programs to encourage external vulnerability discovery.<br><br>Share lessons learned across industry to collectively strengthen defenses.<br><br>Release regular updates to the model that patches security vulnerabilities. |
| IMPLEMENTATION RESOURCES | UC Berkeley CLTC Draft Profile under Measure 2.7 |
| **Produce a "Pre-Systems Card"** | Disclose planned testing, evaluation, and risk management procedures for foundation/frontier models prior to development. |
| BASELINE PRACTICES | Produce a "pre-systems card" detailing quality management plans before research begins.<br><br>Outline intended training data approach, model testing, safety evaluations, responsible AI practices, and development team.<br><br>Submit pre-systems cards to regulatory bodies for review where required.<br><br>Update plans as needed throughout the R&D process.<br><br>*(Note: baseline practices for disclosure practices are still emerging across stakeholders.)* |
| RECOMMENDED PRACTICES | Prepare a written "safety case", explaining why the model is safe enough to develop.<br><br>Share pre-systems cards with independent experts for external review.<br><br>The guidance on Responsible Iteration guides testing and release practices, while pre-systems cards disclose development plans before implementation. |
| **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.<br><br>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.<br><br>Proactively identify and minimize potential sources of bias in training corpora, and adopt techniques to minimize unsafe model behavior.<br><br>Conduct evaluations using cross-disciplinary review teams spanning ethics, security, social science, safety and other relevant domains.<br><br>Address identified risks and adapt deployment plans accordingly based on learnings from pre-deployment evaluations.<br><br>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. |
| **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. |
| BASELINE PRACTICES | Provide controlled access to models for additional evaluative testing by external researchers. External evaluators should be granted sufficient model access, computational resources, and time to conduct effective evaluations prior to deployment.<br><br>Consult independent third parties to audit models following prevailing best practices on methodologies.<br><br>Implement appropriate safeguards to prevent unauthorized access or information leaks via external evaluations.<br><br>Address identified risks and adapt deployment plans accordingly based on learnings from pre-deployment evaluations.<br><br>Maintain documentation of evaluation methods, results, limitations, and steps taken to address identified issues and integrate insights in public reporting per guidance below, except for cases where sharing findings carries sufficient risk of harm.<br><br>*(Note: Enabling robust third-party auditing remains an open challenge requiring ongoing research and attention.)* |
| RECOMMENDED PRACTICES | Pursue diverse external assessment methods including panels and focus groups.<br><br>Collaborate with third parties to support creation of context-specific auditing methodologies focused on evaluating real-world impacts in specific domains and use cases per guidance below. |
| 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.<br><br>Conduct iterative red teaming throughout model development. Continuously evaluate results to identify areas for risk mitigation and improvements, including for planned safeguards.<br><br>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.<br><br>Address identified risks and adapt deployment plans accordingly based on learnings from pre-deployment evaluations.<br><br>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. |
| **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. |
| BASELINE PRACTICES | Publish "key ingredient list" which can include the model's compute, parameters, architecture, training data approach, and model documentation, except for cases where sharing findings carries sufficient risk of harm.<br><br>Disclose details such as performance benchmarks, domains of intended and foreseeable unintended use, risks, limitations, and steps to mitigate risks.<br><br>Disclose details such as on testing methodologies, evaluation criteria, results, limitations, and gaps for any internal and external evaluations conducted prior to release. Integrate relevant insights from responsible iteration practices per guidance above.<br><br>Disclose potential environmental and labor impacts per guidelines below. |
| RECOMMENDED PRACTICES | Collaborate across industry, civil society, and academia to advance public reporting practices weighing transparency with privacy, safety, and other tradeoffs.<br><br>Align disclosures with existing and emerging best practices like Model Cards, System Cards, Datasheets, Fact Sheets, Nutrition Labels, Transparency Notes and Reward Reports.<br><br>Take an incremental approach to transparency disclosures, prioritizing information users rely on to assess capabilities and risks. Progress to more comprehensive disclosures over time as stakeholders gain experience with disclosure practices. |
| **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><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.)* |
| BASELINE PRACTICES | Provide clear documentation to downstream developers covering appropriate uses, limitations, steps to mitigate risks, and safe development practices when building on frontier models.<br><br>Ensure documentation follows prevailing industry standards and accepted best practices for responsible AI development. |
| RECOMMENDED PRACTICES | Collaborate with civil society and downstream developers to advance documentation standards that meet the needs of developers when models are offered through restricted access. This can include gathering inputs on:<br><br>• Safe development checklists for building responsibly on restricted models.<br>• Preferred channels for usage guidance and addressing developer questions, aligned with guidance below on enabling feedback mechanisms. |

| | |
|---|---|
| **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. |
| BASELINE PRACTICES | Embed safety features directly into model architectures, interfaces and integrations<br><br>Publish clear terms of use prohibiting harmful applications and outlining enforcement policies.<br><br>Limit access through approved applications by implementing appropriate identity/eligibility verification requirements to restrict misuse, along with other technical controls like rate limiting and content filtering.<br><br>Maintain processes to regularly re-evaluate technical and procedural controls, monitor their effectiveness including robustness against jailbreaking attempts, and update terms of use as potential misuses evolve. |
| RECOMMENDED PRACTICES | Collaborate across industry and civil society to identify emerging threats requiring new safeguards.<br><br>Provide appropriate transparency into safeguards while protecting integrity.<br><br>Take additional steps to ensure that terms of use are read and understood by users. |

## Post-Deployment

| | |
|---|---|
| **Monitor deployed systems** | Continuously monitor foundation/frontier models post-deployment to identify and address issues, misuse, and societal impacts. |
| BASELINE PRACTICES | Establish ongoing monitoring procedures for deployed models covering areas like performance, fairness, unintended uses, misuses, and risks from combining the model with others.<br><br>Define processes to detect issues and respond appropriately, including notifying partners of significant incidents and considering restricting or retiring a model per guidelines below. |
| RECOMMENDED PRACTICES | Provide transparency into monitoring practices, while protecting user privacy.<br><br>Assess downstream real-world impact of models, for example in collaboration with external researchers.<br><br>Collaborate across industry, civil society, and academia to identify shared challenges and best practices for monitoring. |
| **Implement incident reporting** | Enable timely and responsible reporting of safety incidents to improve collective learning. |
| BASELINE PRACTICES | Implement secure channels aligned with guidance 19 for external stakeholders to report safety incidents or concerns. Also enable internal teams to responsibly report incidents.<br><br>Notify appropriate regulators and partners of critical incidents according to established criteria.<br><br>(Note that baseline practices for incident reporting are still emerging across stakeholders.) |
| RECOMMENDED PRACTICES | Proactively seek external feedback to improve transparency and effectiveness of incident reporting policies and processes.<br><br>Contribute appropriate anonymized data to collaborative incident tracking initiatives to enable identifying systemic issues, while weighing trade offs like privacy, security, and other concerns. |
| **Establish decommissioning policies** | Responsibly retire frontier models from active use based on well-defined criteria and processes. |
| BASELINE PRACTICES | Establish decommissioning procedures and policies including criteria for determining when to restrict, suspend or retire models.<br><br>• Restrict — Limit model use to reduced set of use cases/applications.<br>• Suspend — Temporarily prohibit all model use for remediation.<br>• Retire — Permanently take model out of service. |
| RECOMMENDED PRACTICES | Continue monitoring retired models for downstream impacts and security vulnerabilities per guidance above to prevent unauthorized access and leaks. |

| | |
|---|---|
| **Develop transparency reporting standards** | Collaboratively establish clear transparency reporting standards for disclosing foundation/frontier model usage and policy violations. |
| BASELINE PRACTICES | Participate in collaborative initiatives to align on transparency reporting frameworks and standards with industry, civil society, and academia, as commercial uses evolve. |
| RECOMMENDED PRACTICES | Release periodic transparency reports following established standards, disclosing aggregated usage insights and violation data. Take appropriate measures to ensure transparency reporting protects user privacy and data. |

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

| | |
|---|---|
| **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. |
| BASELINE PRACTICES | Provide sufficient transparency into models and datasets to enable independent assessment and auditing by third parties such as academics and civil society. (Note: Enabling robust third-party auditing remains an open challenge requiring ongoing research and attention).<br><br>Collaborate with third parties to support creation of context-specific auditing methodologies focused on evaluating real-world impacts in specific domains and use cases, beyond base-model evaluations which focus on societal impact evaluations that are not tied to a specific application context. |
| **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.<br><br>Provide or contract with vendors that provide clear instructions for enrichment tasks that are tested for clarity. Enable workers to opt out of tasks.<br><br>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.<br><br>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.<br><br>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. |
| **Conduct human rights due diligence** | Implement comprehensive human rights due diligence methodologies to assess and address the impacts of foundation/frontier models. |
| BASELINE PRACTICES | Establish processes for conducting human rights impact assessments pre-deployment.<br><br>Align with relevant guidance like the UN Guiding Principles on Business and Human Rights, and White House Blueprint for AI Bill of Rights. Proactively assess and address potential impacts on vulnerable communities.<br><br>Continuously improve due diligence processes by collaborating with stakeholders and incorporating community feedback. |
| RECOMMENDED PRACTICES | Publicly disclose identified risks, due diligence methodologies, and measures to address impacts. |

| | |
|---|---|
| **Enable feedback mechanisms across the AI value chain** | Implement inclusive feedback loops across the AI value chain to ethically identify potential harms. |
| BASELINE PRACTICES | Provide clear feedback channels for application developers, consumers, and other direct users. |
| RECOMMENDED PRACTICES | Proactively gather input from indirect stakeholders affected by AI systems through ethical community engagement.<br><br>Establish processes for reviewing feedback and integrating affected user perspectives into development and policy decisions. |
| **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.<br><br>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.<br><br>Incorporate impacts into model development decisions.<br><br>Collaborate across industry, civil society, and academia to advance the measurement of environmental impacts and responsible disclosure practices. |
| **Disclose synthetic content** | Adopt responsible practices for disclosing synthetic media and advance solutions for identifying other synthetic content |
| BASELINE PRACTICES | Provide disclosure mechanisms (both direct disclosure that is viewer or listener facing and indirect disclosure that is embedded) for those creating and distributing synthetic media — content that is not identifiable to the average person and may simulate artifacts, persons, or events.<br><br>Evaluate robustness, ease of manipulation, privacy implications, societal impact, and inherent tradeoffs of different disclosure methods. Provide transparency into assessments and rationale behind final disclosure decisions.<br><br>See Section 2 of PAI's Responsible Practices for Synthetic Media for more information and practices for those building the models for synthetic media. |
| RECOMMENDED PRACTICES | Collaborate across industry, civil society, and academia to advance interoperability and standardization for disclosure of synthetic media.<br><br>Research, develop, and distribute solutions to enable identification and disclosure of synthetic content, including voice and text. |
| **Measure and disclose anticipated severe labor market risks** | Measure and disclose potential severe labor market risks from deployment of foundation/frontier models. |
| BASELINE PRACTICES | Establish clear thresholds for determining when labor market risks are sufficiently severe to warrant disclosure. Consult experts and affected communities in setting disclosure thresholds.<br><br>Conduct assessments to evaluate likely labor market risks and determine their potential severity. Share findings and methodologies used for risk evaluation.<br><br>Regularly review and update severity thresholds as technologies and applications evolve. |
| RECOMMENDED PRACTICES | Collaborate across industry, civil society, academia, and worker organizations to advance the measurement, responsible disclosure practices, and mitigation of severe labor market risks. |