The Cohere Secure AI Frontier Model Framework

cohere

The Cohere Secure AI Frontier Model Framework

V1.0

This document describes Cohere's holistic ***secure AI approach to enabling enterprises to build safe and secure solutions for their customers***. This is the first published version, and it will be updated as we continue to develop new best practices to advance the safety and security of our products.

Introduction

Cohere is the leading security-first enterprise AI company. We build cutting-edge foundation AI models and end-to-end products designed to solve real-world business problems. We partner closely with companies to deliver seamless integration, deep customization, and easy-to-use solutions for their workforce. Our comprehensive range of deployment options offers enterprises the highest levels of data security, privacy, and optionality to deploy across all major cloud providers, private clouds environments, or on-premises.

Our AI solutions are deployed worldwide by our customers and partners in the private and public sectors for applications as varied as customer support, supply chain management, searching and organizing financial information, analyzing legal agreements, answering questions about company policies, and summarizing textual information. These use cases – and many more – are demonstrated across a wide range of sectors and domains, including financial services, telecom, public services, healthcare, manufacturing, retail, and energy. Our models are being used now, at scale, to enable businesses of all sizes to enhance their offerings and better serve their customers.

Enterprises operating in these domains expect and require providers to meet high standards for safety and security, which are determined by existing risk management practices, compliance policies, sector-specific regulations, and broader societal obligations. It is of paramount importance – to us, to our customers, and to the businesses and societies which they serve – that our models are developed and deployed to meet these regulatory, safety, and security needs, today and in the future.

In addition to serving our enterprise customers, we are committed to ensuring that the safety and security of our AI solutions benefit the AI ecosystem and society more broadly. There is a great deal of research and knowledge about the real-world impact that AI already has on society. We play a leading role in forwarding this body of knowledge and implementing practices that minimize harm while benefiting society and businesses.

What does it mean to develop AI solutions for enterprise?
Cohere's business-to-business deployment focus represents a critical and distinct voice within the AI ecosystem. Most AI models, including Cohere's, are deployed through, or combined with, applications, products, or services – in many cases provided by everyday businesses or organizations (enterprise). This means that the way in which models are deployed, the users who have access to them, and the data the models are connected to, matters when it comes to how risks manifest and should be mitigated.

Enterprise AI risks and corresponding mitigation strategies are distinct from AI models integrated into consumer-facing chatbots or smartphone AI assistants that are available to the general public and connected to personal or publicly-available data sources. Examples of how Cohere's AI solutions are used by businesses across various sectors include: identifying patterns in internal reports to improve the safety of production processes in the manufacturing industry; analyzing market data through a custom internal AI workspace designed for the financial sector; and surfacing reliable information across languages from company policies to answer questions from internal and external stakeholders.

With Cohere's focus on secure AI delivering practical yet robust AI solutions for the enterprise market, we prioritize mitigating risks most salient to enterprises. Enterprises are required to satisfy an ever-growing list of standards and requirements, and they consider large language models (LLMs) "safe for use" when they are both *safe*, in that they don't result in harmful outputs, and *secure*, in that systems and data can't be breached. Cohere prioritizes efforts to ensure that these safety and security needs are met. This includes enabling flexible and secure deployment options, such as fully private deployments where customers can run Cohere's AI solutions on-premises or in their own virtual private cloud, where Cohere has no access to customer data or computing environments.

In an enterprise environment, managing AI risks involves protecting both AI models and the broader environments (information and information systems) in which they are deployed from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability. It also involves ensuring our AI products do not cause harm to people, communities, organizations, or wider society.

Our approach to managing AI risks is summarized through the three core principles below:

01

Evidenced Risks

We focus on real-world risks that are known, measured, or observable.

02

Assessed in Context

Risks are identified and assessed in the context of both Cohere's and Cohere's customers' specific use cases and contexts.

03

Holistically Managed

AI risk management is embedded by design into model and system development.

The Cohere Secure AI Frontier Model Framework, outlined in this document, describes how we implement this approach in practice.

Our framework at a glance

The Cohere Secure AI Frontier Model Framework contains five components: (i) Risk Identification; (ii) Risk Assessment and Mitigation; (iii) Risk Assurance Mechanisms; (iv) Transparency; and (v) Research and External Stakeholder Engagement. Risk management practices are implemented in each of these components.

Framework component	Our activities and practices
1. Risk Identification Understanding the type and nature of risks associated with our AI models and systems.	1a Identifying risks arising from model capabilities 1b Identifying risks to our systems 1c Understanding possible harms in context
2. Risk Assessment and Mitigation Applying measures to reduce risk throughout the lifecycle of our models and systems.	2a Defense-in-depth 2b Addressing harms 2c Secure AI by design
3. Risk Assurance Mechanisms Conducting tests and evaluations to verify and demonstrate that AI risks are appropriately mitigated.	3a Evaluations and benchmarking 3b Internal and third-party vulnerability testing 3c Customer-led testing and evaluation
4. Transparency	4a Documenting and sharing our practices

Sharing information about our risk management practices and inviting feedback to ensure continuous improvement.	4b Continuous improvement
5. Research and External Stakeholder Engagement Supporting independent research and engaging externally to advance the science of AI risk management and contribute to industry standards.	5a Support for independent research 5b External engagement

Framework in depth

1) Risk Identification

Risk identification is the first step to understanding the type and nature of risks associated with our AI solutions. We identify risks by first assessing potential risks arising from our models' capabilities and the systems in which they may be deployed. We then assess the *likelihood* and *severity* of potential harms that may arise in enterprise contexts from the identified risks.

1a) Identifying risks arising from model capabilities

Large language models (LLMs) are capable of generating high-quality content, such as text or code, and they can ground their output in context to increase its relevance. However, LLMs also have inherent limitations arising from their architecture, training data, and the nature of the underlying technology. Because LLMs generate content based on statistical patterns, they are limited in the extent to which they can leverage context to inform content generation. Limitations in training data, such as unrepresentative data distributions, historically outdated representations, or an imbalance between harmful patterns and attributes on the one hand and positive patterns and attributes on the other, also impact model capabilities. If these limitations are not mitigated, models can output harmful content, such as hateful or violent content, or child sexual exploitation and abuse material (CSAM).

We therefore focus our secure AI work on risks that have a high likelihood of occurring based on the types of tasks LLMs are highly performant in, as well as the limitations inherent in how these models function. This is what we refer to as “model capabilities.”

We place potential risks arising from LLM capabilities into one of two categories:

1. Risks stemming from **possible malicious use** of foundation AI models, such as generating content to facilitate cybercrime or child sexual exploitation
2. Risks stemming from **possible harmful outputs** in the ordinary, non-malicious use of foundation models, such as outputs that are inaccurate in a way that has a harmful impact on a person or a group

1b) Identifying risks to our systems

Cohere consistently reviews state-of-the-art research and industry practice regarding the risks associated with AI, and uses this to determine our priorities. At Cohere, risks to our systems are identified through a list of continuously-expanding techniques, including:

- Mitigating core vulnerabilities identified by the Open Worldwide Application Security Project (OWASP)
- Internal threat modeling, which includes a review of how our customers interact with and use our models, to proactively identify potential threats and implement specific counter measures before deployment
- Monitoring established and well-researched repositories of security attacks and vulnerabilities for AI, such as the Mitre Altas database

With these methods, Cohere can identify risks such as data poisoning, model theft, inference attacks, injection attacks, and output manipulation.

Moreover, we identify risks across our broader technology stack and environment by performing continuous monitoring of our security controls using automated and manual techniques. Models are developed and deployed in broader computational environments, and effectively managing AI risks requires us to identify, assess, and mitigate information security threats or vulnerabilities that may arise in these environments.

1c) Understanding possible harms in context

The final step to our risk identification process involves understanding the *likelihood* and *severity* of harms that could potentially arise from the risks identified in Section 1a and Section 1b, considering the enterprise contexts in which our AI solutions are deployed.

We categorize possible harms broadly into two categories:

1. **Harm to individual users** (e.g., exposure to content that is hateful or violent)
2. **Societal harm** (e.g., language models that consistently fail for specific demographic groups, or large-scale harms such as violation of data privacy rights resulting from insecure code or malware)

In the chart below, we further detail specific potential harms within each category, and provide an illustrative assessment of their likelihood and severity.

Potential Harm	**Use Case**	**Malicious or Unintentional**	**Likelihood of Harm in Context**	**Severity of Harm in Context**
Outputs that result in a discriminatory outcome[1]	Resume summarization for human resources managers making hiring decisions, using resume data stored in a company's virtual private cloud.	Unintentional	High, there is a large body of research on this potential harm.	High, in the employment context described here, the harm could involve the loss of an employment opportunity for an individual. At scale, this can lead to societal harm for groups of individuals.
Insecure code	Code generation for enterprise developers managing a company's proprietary data within on-premises servers.	Unintentional	Medium to High possibility of a vulnerability being introduced into company code	Medium to High, depending on the nature of the vulnerability introduced and the type of data handled by the company. Severe vulnerabilities can leave companies vulnerable to cyber attacks affecting individuals and society.
Child sexual exploitation and abuse	Prompting an AI writing assistant to generate stories involving child sexual abuse material.	Malicious	Low in enterprise contexts given the controls typically implemented by enterprises on the acceptable use of internal assets.	Very High, as the harm to children can involve extremely severe bodily and psychological harm.

[1] Enterprise customers may be subject to legal obligations not to make decisions that are discriminatory based on legally protected grounds.

Malware code	Prompting a model to generate code that could be used to spread malware in a cybercrime.	Malicious	Low in enterprise contexts given the controls typically implemented by enterprises on the acceptable use of internal assets.	High to Very High, depending on the code generated and how it is misused. Possible harms resulting from cybercrime can involve widespread financial loss, identity theft, psychological harm to victims, etc.

The examples provided above consider the likelihood and severity of potential harms in the enterprise contexts in which Cohere models are deployed. A similar assessment of potential harms from the same models deployed in contexts such as a consumer chatbot would result in a different risk profile.

To effectively manage AI risks, it is important to identify potential risks based on a contextual assessment of model capabilities, the systems in which they will be deployed, and the likelihood and severity of potential harms.

2) Risk Assessment and Mitigation

Once risks have been identified, they must be assessed and mitigated. We assess and mitigate risks across the development lifecycle of Cohere's AI solutions. This is critical to ensuring that the *potential* harms identified above do not materialize into *actual* real-world harms.

For example, if unrepresentative data distributions or toxic data are left unaddressed, this can lead to models outputting content that is legally discriminatory, hateful, or violent, or involves the sexual abuse of children. Similarly, if vulnerabilities such as data poisoning or prompt injection aren't found and mitigated, this could lead to malicious actors being able to exploit models to cause harm by, for example, exposing sensitive information.

In this section, we lay out our approach to holistically managing risk by implementing safeguards throughout the development lifecycle of an AI model.

2a) Defense-in-depth

At Cohere, we recognize that properly securing AI requires going beyond traditional controls. Cohere's security-first culture drives how we work together to design, operate, continuously monitor, and secure both our internal environment (i.e., network, applications, endpoints, data, and personnel) and customer and partner deployments. In addition to applying a holistic approach to security, additional protective measures are woven into the fabric of our models. We cannot secure our models if we do not secure the environment in which we develop them.

Put another way, we deploy a **defense-in-depth** strategy. This means we apply layered controls as part of our overall security management program across all systems and processes – for model development and general day-to-day operations – all of which directly contribute to how we secure our models and our internal environment. We align our program to SOC 2 Type II and other recognized frameworks, and we rigorously monitor the health and performance of our security controls throughout the year, performing real-time corrective action when needed.

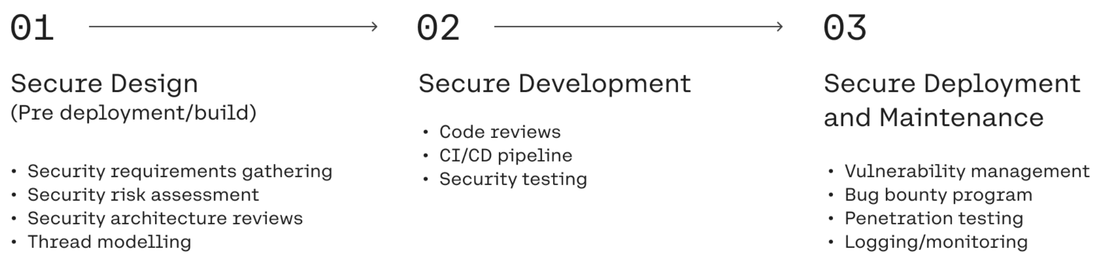
Core security controls

Our core controls across network security, endpoint security, identity and access management, data security, and others are designed to protect Cohere from cyber risks that could expose our models, systems, or sensitive data, such as malware, phishing, denial-of-service, insider threats, and vulnerabilities.

These controls include:

- Advanced perimeter security controls and real-time threat prevention and monitoring
- Secure, risk-based defaults and internal reviews
- Advanced endpoint detection and response across our cloud infrastructure and distributed devices
- Strict access controls, including multifactor authentication, role-based access control, and just-in-time access, across and within our environment to protect against insider and external threats (internal access to unreleased model weights is even more strenuously restricted)
- “Secure Product Lifecycle” controls, including security requirements gathering, security risk assessment, security architecture and product reviews, security threat modeling, security scanning, code reviews, penetration testing, and bug bounty programs

Secure Product Lifecycle Process



01

Secure Design
(Pre deployment/build)

- Security requirements gathering
- Security risk assessment
- Security architecture reviews
- Thread modelling

02

Secure Development

- Code reviews
- CI/CD pipeline
- Security testing

03

Secure Deployment and Maintenance

- Vulnerability management
- Bug bounty program
- Penetration testing
- Logging/monitoring

Deployment-specific controls

Where applicable, we also consider risks within the context of customer deployments. For example, because many of our users start building applications through our application programming interfaces (APIs) before moving to more advanced deployments, we extensively test and secure our APIs. Our API V2 underwent a heavy security design review before we made it available.

For more sophisticated deployments, we focus heavily on container security as our models are bundled via containers. We proactively update our containers with security fixes and also work with customers to provide updated containers if they identify vulnerabilities. Prior to deployment, significant model releases undergo an independent third-party penetration test to validate the security of containers and models.

2b) Addressing harms

We leverage a combination of techniques to address and mitigate the potential harms described in Section 1. Section 2c below outlines how these techniques are applied throughout the development process to iteratively evaluate and improve our models. As noted above, we also continuously research ways to mitigate risk in enterprise use cases, and leverage our findings to improve our approach to secure AI.

More specifically, our harm mitigation practices are focused on achieving the following goals:

- Preventing the generation of harmful outputs in multilingual enterprise use cases
- Adhering to guardrails
- Minimizing over-refusal

These objectives are informed by the priorities of our enterprise customers who expect performant models capable of completing tasks without generating harmful outputs in the languages in which they operate, along with the ability to apply effective, tailored guardrails.

Preventing the generation of harmful outputs involves testing and evaluation techniques to control the types of harmful output described in Section 1, for example, child sexual abuse material (CSAM), targeted violence and hate, outputs that result in discriminatory outcomes for protected groups, or insecure code. Cohere also tests and evaluates output generation in the various languages in which Cohere customers do business.

Guardrails, or what we call Safety Modes, are a feature that allows Cohere customers to exercise more precise control over model outputs. For example, a customer who uses Cohere's model for educational or journalistic purposes may want to enable generation of content about violent historical events in order to respond to requests for information. In that case, they could select Cohere's "contextual" Safety Mode as appropriate for their use cases. In contrast, a customer who provides Cohere models to their large workforce may want to ensure that all employees use the models within prescribed, lower-risk limits. In this case, they may want to set guardrails to "strict" Safety Mode in order to prohibit profanity, explicit content, and violence. Beyond simply offering these features, Cohere conducts various evaluations to ensure that models actually adhere to these guardrails.

Over-refusal refers to when a model misinterprets a safe prompt as a harmful one and does not comply with a user's request. Over-refusal can cause performance problems. But if a model refuses to complete tasks in a way that is detrimental to a specific group or individual – if it more frequently refuses to summarize non-English resumes, for example – it can also be a potential harm in itself. We aim to minimize such false-positive cases.

Cohere's models, their training data, and the guardrails within which they operate are dynamically updated throughout the development process to achieve the three harm mitigation objectives described above.

Where Cohere has direct visibility into the use of its models during deployment, we use that visibility to monitor for malicious attempts to prompt our models for harmful outputs, revoking access from accounts that abuse our systems. **Cohere partners closely with customers who deploy Cohere's AI solutions privately or on third-party managed platforms to ensure that they understand and recognize their responsibility for implementing appropriate monitoring controls during deployment.**

2c) Secure AI by design

The following table provides an overview of how the risk assessment and mitigation approaches described above are embedded throughout our models' lifecycle, drawing upon long-standing approaches to secure product development and emerging best practices and research for AI risk mitigation.

Stage	**Key Risks**	**Key Mitigations We Apply**
Data acquisition and preparation Data needs are analyzed and planned. Data is then collected, selected, cleaned, analyzed, and processed. Synthetic data is generated and validated. Once datasets are finalized, they are ingested to train a model. Data acquisition and preparation also functions as a circular process as new or updated data needs are identified and addressed throughout the lifecycle.	• Data poisoning • Supply chain vulnerabilities • Model theft • Insecure plugin design • Unrepresentative data distributions • Imbalance of data with harmful patterns and attributes vs. positive patterns and attributes • Historically outdated representations in data • Inaccurate proxies when used to measure representativeness or imbalances	• Detailed data lineage controls, including tracking the source, pre-processing steps, storage location, and access permissions • Supply chain controls for any third parties (e.g., data vendors or third-party data annotation) • Traditional just-in-time access controls, robust authentication, zero-trust rules, etc. • Data pre-processing (including cleaning, analysis, selection, etc.) • Re-sampling, re-weighting, and re-balancing datasets to reduce identified representation issues or imbalances

Training, evaluations and testing Models are trained to perform across a range of tasks. Evaluations and testing for general performance, safety, and security occur throughout the lifecycle and include red teaming, evaluation with academic and industry benchmarks, and internal bespoke evaluations. Algorithms and training data are adjusted iteratively based on evaluation and testing results.	• Data poisoning • Data leakage • Model theft • Adversarial attacks • Evaluation criteria and data are not representative of a population • Disparate performance in different cases results in disproportionate impact on certain populations • Models and data are fit for an aggregated, dominant population but sub-optimal for sub-groups within the population	• Multi-disciplinary red teaming • Independent third-party security testing, e.g., penetration testing • Continuous monitoring to detect anomalies and security issues • Multi-disciplinary red teaming • Consultation of domain experts • Multi-faceted evaluations, including standard benchmarks and proprietary evaluations based on identified possible harms and harm reduction objectives • User research of local language and cultural contexts
Deployment and maintenance Models are deployed and subject to ongoing monitoring controls.	• Prompt injection • Insecure output handling • Model denial of service • Excessive agency • Sensitive information disclosure • Misuse • Unexpected post-deployment usage patterns that were not accounted for and result in unmitigated risk	• Blocklists, custom classifiers, and prompt injection guard filters, and human review to detect and intercept attempts to create unsafe outputs • Specific mitigations applied based on deployment type, e.g., isolated customer environments with focus on remediating security vulnerabilities that coexist between traditional application security and AI security • Security Information and Event Management (SIEM) system leveraging heuristics and advanced detection capabilities to identify potential threats • “Air-gapped" safeguards to prevent lateral movement and unintended network calls across environments and kernel-based LLMs to prevent the leaking of

		shared memories or buffers that could expose sensitive data • Blocklists • Safety classifiers and human review to detect and intercept attempts to create unsafe outputs • Human-interpretable explanation of outputs • User research and customer feedback analysis
Improvement and further fine-tuning User feedback and user research are used for continuous incremental improvements. Fine-tuning may be leveraged by a deployer or by a developer to improve performance in a specific area.	• Prompt injection • Insecure input/output handling • Model denial of service • Excessive agency • Sensitive information disclosure • Adversarial attacks • Evaluation criteria and data are not representative of a population • Model design choices amplify performance disparity across different examples in the data	• Responsible Disclosure Policy to incent third-party security vulnerability discovery • Specific mitigations applied based on deployment type, e.g., isolated customer environments with focus on remediating security vulnerabilities that coexist between traditional application security and AI security • Continuous evaluation and user research • Programs to incentivize research, including research grants and participation in external independent research efforts. • Multi-disciplinary red teaming

3) Risk Assurance Mechanisms

Assurance mechanisms are measures that provide confidence that risk mitigations are effective and expected risk mitigation objectives have been met. Assurance mechanisms are vital for enterprise businesses adopting AI technologies as they provide necessary safeguards, build trust, ensure compliance, and facilitate continuous improvement.

One approach to risk assurance in the AI industry is focused on risks described as catastrophic or severe, such as capabilities related to radiological and nuclear weapons, autonomy, and self-replication. In this context, thresholds relating to these potential catastrophic risks are developed, and the approach described in safety frameworks is designed to assess risks that are speculated to arise when models attain specific capabilities, such as the ability to perform autonomous research or facilitate biorisk. The models are then deemed to present “unacceptable” levels of risk when certain capability levels are attained.

While it is important to consider long-term, potential future risks associated with LLMs and the systems in which they are deployed, studies regarding the likelihood of these capabilities arising and leading to real-world harm are limited in their methodological maturity and transparency, often lacking clear theoretical threat models or developed empirical methods due to their nascency. For example, existing research into how LLMs may increase biorisks fails to account for entire risk chains beyond access to information, and does not systematically compare LLMs to other information access tools, such as the internet. More work is needed to develop methods for assessing these types of threats more reliably.

Cohere’s approach to risk assurance, and to determining when models and systems are sufficiently safe and secure to be made available to our customers, is focused on risks that are known, measurable, or observable today. Assurance mechanisms applied to provide confidence that risk mitigations are effective, as outlined below, are similarly focused on assurance measures that can be measured or observed based on today’s state-of-the-art practices. Some examples include our evaluations and testing, network and API penetration tests, and robust management of container vulnerability in private deployments.

Cohere’s approach to assurance mechanisms is dynamic, much like the approach to risk assessment and mitigation outlined above, and entails multiple layers of assurance involving both internal and external parties. This approach meets the needs of enterprises with robust risk environments like financial services and public sector.

Cohere’s approach to risk assurance also includes clear, final points of evaluation to validate that a model is safe, secure, and ready to be made available to our customers.

The final authority to determine if our products are safe, secure, and ready to be made available to our customers is delegated by Cohere’s CEO to Cohere’s Chief Scientist. This decision is made on the basis of final, multi-faceted evaluations and testing. In addition, customers deploying Cohere solutions may have additional tests or evaluations they wish to conduct prior to launching a product or system that integrates Cohere models or systems. Cohere provides necessary support to customers to meet any such additional thresholds.

3a) Evaluations and benchmarking

As described above, Cohere conducts evaluations throughout the model development cycle, using both internal and external evaluation benchmarks. This ensures that our models are not only high-performing, but also safe and reliable. When a model is nearing launch, the modeling function at Cohere conducts a comprehensive final evaluation, assessing both performance and safety metrics. This critical step ensures that our models meet the highest standards before deployment. For example, we evaluate new model versions with industry-standard benchmarks like BOLD (Biases in Open-ended Language Generation) and publish the results.

Any performance regressions identified during any such testing or evaluations, including the final pre-deployment evaluation, are investigated and mitigated before deployment. We consider models safe and secure to launch when our evaluations and tests demonstrate no significant regressions compared to our previously launched model versions, so that performance and security is maintained or improved for every new significant model version. This is Cohere's bright line for determining when a model is "acceptable" from a risk management perspective and ready to be launched.

At Cohere, we prioritize secure AI through a rigorous evaluation process. When a model is nearing launch, the modeling organization conducts a comprehensive final evaluation, assessing both performance and safety metrics. This critical step ensures that our models meet the highest standards before deployment.

3b) Internal and third-party vulnerability testing

Prior to major model releases, Cohere also performs robust vulnerability management testing, including independent third-party penetration testing of model containers, and vulnerability patching and mitigation, to ensure that models are secure enough to launch.

Cohere conducts multidisciplinary red teaming during both the model development phase and post-launch. These red teaming exercises may include independent external parties, such as NIST and Humane Intelligence, and are conducted based on realistic use cases to attempt to break the model's ability to fulfill alignment on risk mitigation goals in order to elicit information about areas of improvement. Results are tracked over time to identify any performance regressions that similarly are mitigated prior to model deployment. For example, Cohere conducted red-teaming on robustness to uncover weaknesses in the model's ability to follow safety instructions even when the specific language used in instructions was changed. The data produced from this exercise was used to develop a robustness evaluation, which was then run on subsequent model versions to provide confidence that the model would be resilient to variability in real-life usage.

3c) Customer-led testing and evaluation

Considering Cohere's focus on serving enterprise customers, assurance mechanisms may also vary by specific deployment and customer context. Cohere provides targeted guidance to customers to assist them in implementing appropriate AI risk mitigation measures for their own AI-driven products and services. This includes Cohere's AI Security Guide and Enterprise Guide to AI Safety. Where requested, Cohere also works in partnership with its customers to conduct supplementary assurance evaluations or testing as needed. In this way, the analysis of whether a model is "acceptable" from a risk management perspective must be adapted to the customer context, and must be able to adapt to new requirements or needs that emerge post-deployment. Assurance here means working with our customers to ensure that our models and systems conform to their risk management obligations and standards.

4) Transparency

To provide information to those outside Cohere – customers, government agencies, and the wider public – and to contribute to the development of best practices, we record and make available information about our risk management practices.

4a) Documenting and sharing our practices

Documentation is a key aspect of our accountability to our customers, partners, relevant government agencies, and the wider public. To promote transparency about our practices, we:

- Publish documentation regarding our models' capabilities, evaluation results, configurable secure AI features, and model limitations for developers to safely and securely build AI systems using Cohere solutions. This includes model documentation, such as Cohere's Usage Policy and Model Cards, and technical guides, such as Cohere's LLM University.
- Are publishing this Framework to share our approach on AI risk management for secure AI.

- Offer insights into our data management, security measures, and compliance through our Trust Center.
- Provide guidance to our Customers on how they can manage AI risk for their use cases with our AI Security Guide and Enterprise Guide to AI Safety.

4b) Continuous improvement

We are constantly improving our practices to better mitigate AI risks. To identify areas of improvement, we engage with our customers, partners, and the wider public by:

- Incentivizing third-party vulnerability discovery via clear protections for legitimate research practices in our Responsible Disclosure Policy
- Conducting user research to understand what challenges and risks can be expected in enterprise use cases
- Engaging with the broader community via dedicated user forums

We are constantly internally evaluating the tools, techniques, and products we use across the deployment lifecycle to identify and mitigate AI risks as the industry evolves and new techniques become available.

5) Research and External Stakeholder Engagement

5a) Support for independent research

The field of AI risk management extends beyond individual company practices to encompass a wide ecosystem of researchers across industry, government, civil society, and academia. This collective effort contributes to advancing research into AI risks, as well as developing techniques and best practices that can be adopted by model developers.

Cohere actively contributes to this ecosystem, in large part through our open science research lab, Cohere For AI (C4AI). In addition to conducting fundamental research – on topics including, but not limited to AI risks - C4AI supports a community of over 3000 researchers around the world to connect, collaborate, and share research with one another, and expand how AI research is done and by whom. Additionally, C4AI directly supports independent researcher efforts through a Research Grant Program, which provides researchers in nonprofit, public sector, or academic settings access to use Cohere's models for research or public interest projects via subsidized access to our API.

With respect to AI security, Cohere's bug bounty program provides monetary incentives to promote and advance the understanding of AI security. Cohere also actively contributes to thought leadership initiatives in collaboration with industry peers to improve the overall security of the AI value chain.

5b) External engagement

Cohere is committed to building a responsible, safe, and secure AI ecosystem, and actively engages with external actors to continuously improve our own practices, as well as to advance the state-of-the art on AI risk management.

In particular, Cohere contributes to the development of critical guidance and industry standards with organisations such as:

1. OWASP Top 10 for Large Language Models and Generative AI
2. CoSAI (Coalition for Secure AI) — founding member
3. CSA (Cloud Security Alliance)
4. ML Commons

Cohere also engages in cooperation with international AI Safety Institutes and external researchers to advance the scientific understanding of AI risks, for example by submitting our public models for inclusion on public benchmarks and red teaming exercises.

We wish to thank researchers at the Centre for Democracy and Technology and the Ada Lovelace Institute for their expertise in reviewing and providing feedback on the Cohere Secure AI Frontier Model Framework.