

OECD AI Transparency Report

Organization: Cohere (CAN)

Reporting Period: Q3 2026

Published: October 2, 2026

Section 0 – Your role in the AI system lifecycle

Responding as (please select all that apply):

- Model developer (/provider)
- Application developer (/provider)

If any of the fields above are selected, please indicate your organization size:

Large enterprise: 250 or more employees

Section 1 – Risk Identification and Evaluation

1. How does your organization define and/or classify different types of risks related to AI?¹

- Maintain an internal taxonomy, classification system, and/or set of risk area definitions
- Define enhanced safety measures for specific dangerous capability thresholds, that, once reached, require safeguards (e.g. capability threshold for dangerous Chemical, Biological, Radiological and Nuclear (CBRN), cybersecurity attacks, dangerous advanced autonomy capabilities, or other capabilities).
- Align with specific external frameworks, standards, reports, and/or guidance documents such as taxonomies, classification systems, and/or sets of definitions for risk areas (e.g. the NIST AI Risk Management Framework, the ISO/IEC 42001, the OECD AI Principles, the EU General-Purpose AI Code of Practice, the International AI Safety Report). (please specify): A range of frameworks and guidance documents, including: ISO 42001; EU Code of Practice for General-Purpose AI; EU Code of Practice for Transparency of AI-Generated Content
- Consult with external stakeholders (e.g., from academia, industry groups, civil society, or standard development organisations) when developing or updating risk taxonomies

How does your organization define thresholds at which severe risks, unless adequately mitigated, may be considered "unreasonable"? Please describe the framework(s) most relevant to your organization's role and context. If your organisation does not use formal thresholds, you may describe instead the policies, decision criteria, or governance processes you apply when assessing severe AI risks. You may also describe:

- Specific examples of contexts in which your models or systems could, in your assessment, pose unreasonable or intolerable risk;
- How mitigations are graduated by capability or propensity level;
- How your organisation considers the specificities of different AI technologies, such as generative AI and agentic AI, including any specific risk assessment or governance mechanisms applied to agentic AI capabilities.

Cohere's approach to identifying and managing severe risks is detailed in our Secure AI Frontier Model Framework, available at: <https://trustcenter.cohere.com/resources-details>

At Cohere, AI risk management is embedded throughout the model and system development lifecycle. New models are considered ready for launch if our performance and safety evaluations show no significant regressions compared to previous versions – thus always driving the standard for safety higher with each iteration. Our framework focuses on known, measurable risks rather than speculative threats. We prioritize enterprise-specific risks where harmful outputs could cause actual harm in business contexts.

2. Both before and after deployment, what practices does your organization use to identify and evaluate vulnerabilities, incidents, emerging risks and misuse throughout the AI system lifecycle?

- Internal red teaming³
- Independent third-party red teaming
- Adversarial testing to detect technical vulnerabilities
- Publicly available benchmarks and measurement tools
- Internal benchmarks, standard metrics (e.g., frequency, severity scores)
- Structured techniques to explore the behavior or performance of a model and/or application, such as model elicitation or bottleneck assessments
- Qualitative assessments (e.g., expert judgment, narrative scoring)
- Specific risk identification practices for vulnerable groups
- Impact assessments such as personal data protection assessment and fundamental rights impact assessments
- Internal model, system, or application-independent research initiatives
- External model, system, or application-independent research initiatives
- Vulnerability reports
- Incident reports, including reports shared by other organizations
- Automated monitoring / remote status reporting (e.g., telemetry pipelines, automated alerts when risk thresholds are breached)
- Specific risk identification practices for agentic AI capabilities, e.g., autonomous action-taking, tool use, multi-step planning (please specify): Tool use, multi-step planning, task-based and context-based assessments.
- Multilingual and cross-cultural evaluation, including collaboration with local ecosystems (e.g., national AI institutes, language-specific evaluation consortia) for under-represented languages and cultural contexts (please specify): Cohere's multilingual modelling team and foundational open science research team (Cohere Labs) have pioneered several multilingual and multicultural research and evaluation projects, with multiple specialised multilingual models and several datasets published. These include: Aya - an international open science initiative to develop datasets and models that serve multiple languages, including many which are underrepresented. See <https://cohere.com/research/aya> -- Command A Translate – an industry-leading translation model for accurate, context-aware multilingual communication across global customer, content, and localization workflows. <https://cohere.com/blog/command-a-translate>

-- Transcribe Arabic - the world's most accurate open-source model for Arabic speech recognition. See <https://cohere.com/blog/transcribe-arabic> -- Global MMLU - an evaluation covering 42 languages to assess cultural and language coverage, accuracy, biases. <https://cohere.com/research/globalmmlu> -- More than 30 research papers – including datasets - published since 2024 which focus on multilingual and multicultural AI performance and safety. See <https://cohere.com/research/papers>

Are any national AI safety / security institutes of the AISI network involved?

Yes

Please briefly describe your organization's decision-making process for deploying an AI model or system including: who has decision authority; what evidence and criteria are required (e.g., performance, safety/security, robustness, fairness, privacy, compliance, independent external testing, evaluation, validation and verification (TEVV) or assurance); contextual caveats or limitations required in interpreting risk evaluation metric results; what approvals/sign-offs are needed; and what safeguards (e.g., mitigations, monitoring, or phased release) must be in place to move into deployment.

Our decision-making process when releasing our new models is detailed in Answer 1A and in our Secure AI Frontier Model Framework, available at: <https://trustcenter.cohere.com/resources-details>. We review safety and performance evaluation results and assess for parity or improvements compared to previous release, to ensure each new model release improves the baseline for safety and capability simultaneously.

Following deployment, how do you use incident reports — including those shared by other organizations — to identify emerging issues, update risk assessments, and trigger corrective actions.

Cohere receives post-deployment incident signals through responsible disclosure and bug bounty channels for customers and researchers. Externally-sourced intelligence comes from open source resources such as MITRE ATLAS, OWASP, industry bodies including Coalition for Secure AI, ML Commons, government security agencies and research institutes, and other threat intelligence sources. We regularly review our risk methodology and controls to keep them aligned with evolving industry practice.

3. Does your organization have mechanisms to receive and/or share reports of risks, incidents or vulnerabilities by/with third parties?

Yes, both receiving and sharing

If yes, please indicate which stakeholders reporting mechanisms may be accessible to, recognising specifics will depend on the nature of the report and risk level. For each, indicate whether access is on a standing basis (Always), activated based on context (Ad hoc/as relevant), or not currently accessible.

	Always	Ad hoc / as relevant	Not currently accessible
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities (e.g., public reporting offices for security incidents or data breaches)		☐	

	Always	Ad hoc / as relevant	Not currently accessible
Supervisory authorities (e.g., data protection authorities)		☐	
Industry partners or industry fora		☐	
Users (end users, enterprise customers, developers using APIs)		☐	
Business customers / deploying organizations (if different from “users”)	☐		
Supply-chain partners (cloud, data providers, etc.)		☐	
Academia / researchers		☐	
Independent auditors		☐	
Standards bodies / multi-stakeholder initiatives (OECD.AI, ISO/IEC, IEEE, etc.)		☐	
Specific civil society organizations		☐	
Media / general public (people or communities indirectly impacted)		☐	
Other			

Please describe how your organisation determines when and how reporting mechanisms are made accessible to different stakeholders (e.g. based on legal or regulatory requirements, internal policies, or risk level). Please also provide public URLs or details on how to submit reports. If none, write “Not applicable”

Cohere maintains distinct, purpose-built reporting channels for varying groups, with the route determined by who is reporting and what type of concern is involved. Safety, abuse and vulnerability reporting is fully public and unauthenticated so that anyone can raise a concern; internally, any employee may report a suspected security incident or other concern through designated channels; customers and security researchers are provided a dedicated responsible disclosure channel e.g. via email inboxes that route directly to our security and safety teams. Vulnerability reports can be submitted through Cohere's bug bounty program. More information and reporting channel details is available at <https://trustcenter.cohere.com/>.

How does your organization share relevant information as appropriate?

- Information-sharing partnerships (with confidentiality agreements as appropriate)
- Reporting to regulatory authorities as relevant

Please describe what type of information is shared.

We work with third-party domain experts, and share relevant incident information with them under confidentiality agreements. This information is often sensitive - access to it could enable misuse or

attacks on our systems or those of others' - hence why it is not shared publicly.

4. If your organization has programs to incentivize risk(s) disclosure, incidents and systemic vulnerabilities, please share what type of programs you use?

- Paid rewards ("bug bounty")
- Vulnerability Disclosure Programs (VDPs) with safe harbor / liability protection
- Early access and collaboration incentives
- Research grants, contracts, and hiring red teamers
- Internal incentives and culture

Please provide public URLs and/or details on your incentive program(s). If none, write "Not applicable."

Information is provided in our Responsible Disclosure Policy, available in our Trust Center
<https://trustcenter.cohere.com/>

5. Does your organization use or contribute to the development of international technical standards, best practices or tools for the identification, assessment, and evaluation of risks?

Contributes and uses

If contributes or uses, please select all that apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001, ISO/IEC 23894, ISO/IEC 42005)
- NIST AI Risk Management Framework (AI RMF)
- Industry-led best practices or codes of conduct
- General Purpose AI Code of Practice
- Other international or regional standards (please specify): EU AI Code of Practice for Transparency of AI Generated Content

6. Does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks that could have broad and potentially cascading socio-economic impacts? Please indicate which stakeholders you collaborate with, and whether collaboration is on a standing basis (Always-on), activated based on context (Ad hoc/as relevant), or there is no current collaboration.

	Always-on	Ad hoc / as relevant	No current collaboration
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities / regulators / AI safety institutes / other public sector bodies and inter-governmental coordination structures		☐	
Industry partners or industry fora	☐		

	Always-on	Ad hoc / as relevant	No current collaboration
Users (end users, enterprise customers, developers using APIs)	☐		
Business customers / deploying organizations (if different from “users”)	☐		
Supply-chain partners (cloud, data providers, etc.)		☐	
Academia / researchers		☐	
Independent auditors		☐	
Standards bodies / multi-stakeholder initiatives (OECD.AI; ISO/IEC; IEEE; AIID etc.)		☐	
Independent civil society organizations (e.g., for evaluations, audits, or policy review)		☐	
Media / general public (people or communities indirectly impacted)		☐	
Other(s) (please add any other relevant stakeholder groups):			
Not applicable			

Please describe examples of how your organization collaborates. Examples may include: participating in multi-stakeholder or industry initiatives; partnering with academia, civil society, or the public sector; establishing internal or joint governance structures with external experts or in partnership with other efforts to share knowledge and jointly address risks.

Cohere participates in several industry coalitions, engages with government working groups and frameworks, and partners with academia and non-profits in a range of ways.

Industry coalitions include: OWASP; ML Commons; Coalition for Secure AI; MITRE ATLAS; ISO.

Government collaborations include: partnerships and engagements with AI security and safety institutes, including an MoU with the Canadian AISI and as members of the NIST AI Consortium; members of Singapore’s AI Verify; participating in the Hiroshima Friends Group; and as members of signatory working groups and task forces for the EU’s AI Act Codes of Practice.

Academic and non-profit partnerships include: partnerships with the Universities of Toronto and Waterloo; research collaborations with academics around the world, led by our technical staff and via our open science research unit, Cohere Labs; research grants which provide compute support to enable research and development as well as external inspection of our models.

Any further comments and implementation documentation.

No answer provided

Section 2 – Risk Management and Information Security

7. Across the AI system lifecycle, what steps does your organization take to manage risks and vulnerabilities? Please indicate which specific practices, tools or metrics your organization uses at each lifecycle stage. You may refer to tools or metrics present in the OECD.AI Catalogue of Tools and Metrics or suggest additions. Respondents may complete only the lifecycle stages relevant to their organisation's role. Stages that do not apply may be left blank.

Plan and design

Collect and process data

Build, integrate and interpret model

Verify and validate

Deploy

Operate and monitor

Please provide a consolidated description of your organisation's approach to risk management across the AI system lifecycle, including how the practices identified above operate together and any cross-cutting governance, review or escalation processes. Please also indicate how often these practices are reviewed and what types of events trigger reviews and updates. Where relevant, please distinguish between different types of AI systems (e.g. generative or agentic AI).

We apply a range of measures across the AI lifecycle

- Plan and design - Measures include risk identification and an assessment of needs for the development process, with a focus on enterprise-relevant capabilities.
- Collect and process data - Measures include Data de-duplication, filtering, and quality controls.
- Build, integrate and interpret model - Measures include decentralized training with self-refinement algorithms and model merging techniques; reward models for evaluation; and parameter merging methods.
- Verify and validate - Measures include Extensive benchmarking across enterprise tasks and public evaluations; red teaming.
- Deploy - Measures include privacy and security-preserving deployment options including private and on-premises options.
- Operate and monitor - Measures include, as applicable to the deployment context, isolated customer environments, vulnerability management, and security information and event management systems.

Through these measures and across the AI lifecycle, we manage risk according to three principles: evidenced risks, contextual assessment, and holistic management by design. These apply across five interconnected components: Risk Identification, Risk Assessment and Mitigation, Risk Assurance, Transparency, and Research Engagement.

Additionally, governance & Review operates across:

- Escalation: Final authority over model deployments lies with Executive Leaders for model safety decisions
- Reviews: Annual framework assessments or triggered by incidents/capability changes
- Documentation: Maintained with version control and changelogs

For more information see our Secure AI Frontier Model Framework, available at:

<https://trustcenter.cohere.com/resources-details> and our Command A technical report, available at

<https://arxiv.org/pdf/2504.00698>

Please describe any measures used to ensure data quality and fairness across the data collection and curation stage (e.g., dataset audits, filtering training data to remove harmful content or personal information, fine-tuning on curated data, use of moderation tools, assessment of limitations to model generalisability).

Cohere implements industry best practices and techniques across the data collection and curation stage. Specifically, we implement techniques to support data quality and suitability for training, including deduplication, filtering for toxic, harmful or otherwise unsuitable content, and quality filtering. Additionally, we employ specialized data annotation processes with multiple reviewers and iterative refinement for general and specialised safety topics - such as ensuring safety across multilingual contexts - and we curate data collection for its usefulness for enterprise relevant-tasks, such as coding, tool use, agentic capabilities, multilingual tasks, document understanding, instruction-following tasks, and long-context tasks.

8. What security measures are used in testing AI systems?

- Use of synthetic, anonymized, or dummy data, as appropriate
- Use of “AI-safe” test-data, as appropriate
- Audit logging and monitoring
- Disabling, restricting or simulating high-impact capabilities during testing
- Access controls, including roles-based permissions
- Human-in-the-loop controls for higher-risk test cases
- Rate limits and resource caps
- Controlled security vulnerability testing
- Containment measures for prompt-injection and jailbreak attempts
- Controlled environments for red-teaming and adversarial testing
- Isolation of staging environments (from production / operational environments)
- Restrictions on outbound Internet or API access, including a list of permitted domains or APIs
- Network isolation from external access
- Locally hosted testing environments
- Isolated and secure test environments

8.A. Please indicate, if relevant, the team or role responsible for overseeing secure testing and whether this responsibility is shared across multiple teams and/or functions.

Cohere's Security team is responsible for secure testing, working closely with Core and Safety modelling teams.

9. How does your organization protect intellectual property, including copyright-protected content?

9.A. Safeguards that promote lawful data use and transparency in AI training:

- Maintain a copyright policy that addresses data sourcing, content provenance, safeguards, and organizational roles and responsibilities
- Publishing "model cards" or contextual information on training data and model limitations
- Apply provenance controls to guard against training data being sourced from websites or repositories that have been determined to engage in large-scale piracy
- Refrain from the unlawful circumvention of technological protection measures in data sourcing
- Maintain policies or practices to respect copyright holders' preferences through, where appropriate/applicable, mechanisms like Robots.txt "do not train" tags or other machine-readable protocols developed through voluntary, industry-led standards processes
- Conduct data cleaning, qualification, and/or tokenization activities as part of a transformative data preparation process
- Consult and/or retain an intellectual property attorney to protect, enforce, and comply with, intellectual property rights.

9.B. Safeguards that reduce the risk of dissemination of outputs that may infringe copyright:

- Designating clear points of contact so creators or rights holders can raise concerns about infringing outputs
- Other(s) (specify): Cohere implements additional technical measures throughout the development lifecycle (e.g. data filtering, regularisation techniques, data curation, private deployments, enterprise-focused model capabilities, and grounding of outputs on customer-provided, enterprise data.

10. How do you protect privacy and prevent your systems from disclosing confidential or sensitive data across the AI lifecycle?

- Make information about data privacy, practices, and controls accessible to a diverse set of stakeholders
- Use security controls, including encryption at rest and in transit
- Provide organizational customers with controls to configure and implement privacy choices (e.g., data retention, logging, access, or processing settings)
- Apply Privacy-Enhancing Technologies (PETs) where applicable (e.g. differential privacy, federated learning, synthetic data generation, homomorphic encryption)

Describe key privacy measures and data-protection standards applied.

Cohere implements comprehensive privacy measures across the AI lifecycle as described in its Model Training Privacy Notice. Available at: <https://trustcenter.cohere.com/resources-details>

11. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems across the AI system lifecycle? Respondents may complete the checkbox subsections in 11.A–11.D or describe their overall approach under 11.E.

11.A. Governance, policy and assessments

- Conduct regular cybersecurity risk and vulnerability assessments tailored to AI models or systems
- Perform threat modeling that includes AI-specific risks (e.g., data poisoning, prompt injection)
- Maintain or update security policies aligned with recognized standards (e.g., NIST CSF, ISO/IEC 27001, SOC 2, CIS Critical Security Controls), and review them regularly in response to emerging risks or regulatory guidance
- Maintain vulnerability management processes (scanning, channels for external reporting, patching, remediation)
- Implement incident response, disaster recovery, or business continuity plans, including escalation procedures for high-severity incidents
- Train staff on cybersecurity best practices and AI-specific threats, including how to recognize and respond to misuse or unintended outcomes of advanced AI systems
- Integrate cybersecurity AI risks into broader cybersecurity risk management activities
- Regularly review security measures and update them in response to emerging risks, incidents, or major releases

11.B. Technical and operational controls

- Conduct security testing focused on AI-specific risks (e.g., red-teaming, adversarial testing, stress tests)
- Conduct secure AI development and/or deployment practices
- Monitor system activity and log access to detect anomalies or misuse
- Restrict physical access to servers, GPUs, or devices hosting AI models, systems, and/or applications
- Encrypt AI assets (datasets, models, configurations) in storage and in transit as appropriate
- Apply multi-factor authentication (MFA) and role-based access controls for advanced AI systems

11.C. Protection of sensitive or high value AI assets

- Limit access to proprietary AI assets based on role or need-to-know
- Store model weights, datasets, and algorithms in secure, access-controlled environments
- Use hardware security modules (HSMs) or trusted execution environments (TEEs) where feasible
- Track and log access to sensitive AI resources
- Require NDAs or confidentiality agreements for staff and contractors with access

11.D. Insider risk protection detections

- Implement insider-risk escalation or response procedures
- Provide staff training on cybersecurity best practices and AI-specific threats
- Monitor for unauthorized internal access or suspicious activity
- Conduct pre-employment screening or role-sensitive vetting

11.E. Please describe the main security measures your organization applies and indicate which are specific to particularly sensitive AI assets, where relevant.

At Cohere, we recognize that properly securing AI requires going beyond traditional controls. Cohere's security-first culture drives how we work to design, operate, 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 through 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.

12. What is your organization's vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?

- Recognition of external finders of confirmed risks or vulnerabilities
- Reporting to regulators / authorities when relevant
- Vulnerability disclosure program that is aligned with internationally recognized standards, e.g. ISO/IEC 30111 and 29147.
- Collaborative vulnerability sharing (e.g., with partners, industry groups)
- Scanning for and mitigation of similar risks or vulnerabilities
- Testing of patches or other mitigations
- Development of patches or other mitigations for confirmed risks or vulnerabilities
- Investigation of reported or identified risks and vulnerabilities
- Bug bounty or incentive programs for external reporters
- Internal vulnerability disclosure receipt and tracking processes
- Defined escalation criteria and follow-on actions for identified vulnerabilities (e.g., for external reporting or information sharing)

13. What post-deployment practices does your organization use to monitor and respond to vulnerabilities, incidents, emerging risks and misuse of advanced AI systems?

- Conduct post-deployment monitoring
- Perform root-cause analysis and apply lessons learned to improve future development and/or deployment practices
- Monitoring and emergency controls specific to agentic AI systems (e.g., ability to interrupt autonomous action sequences, override tool-use permissions, or revoke delegated authority)

14. Do you develop or deploy agentic AI systems (systems that can autonomously plan/act or use tools)?⁴

Yes

14.A. If yes, please indicate which categories of controls your organization applies. Please select all.

- Constraints on the agent's action space (e.g., tool-use permissions, environment sandboxing, allowlists/blocklists of resources, scope limits per task)
- Human oversight or interruptibility mechanisms (approval requirements for sensitive actions, escalation paths, interruption interfaces)
- Monitoring (anomaly detection, multi-agent interaction monitoring)
- Logging of agent actions

Please provide any further comments and implementation documentation.

No answer provided

Section 3 – Transparency Reporting on Advanced AI Systems

15. Does your organization publish public reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems? Are these clear and understandable by non-technical audiences?

Yes

15.A. If yes, which of the following information is included in your organization's publicly available documentation for any of your AI products?

- Risk management policies
- Summary of results from evaluations of safety, security, and societal risks including potential adverse impacts on individuals or groups (e.g. impacts on human rights, privacy or differential outcomes across user groups).
- Summary of results from red-teaming and other testing (e.g., adversarial evaluations, stress tests, and other pre-deployment safety assessments).
- Capabilities, behaviours, and limitations, including domains of appropriate and inappropriate use or indicative thresholds associated with systemic risks.
- Technical documentation and instructions for use if relevant.
- Plain-language or non-technical summaries

Please provide links or references showing where this information is made public.

Our approach to and policies for Risk Management - including our Secure AI Model Framework - can be found on our Trust Center <https://trustcenter.cohere.com/>

Technical documentation, plain language summaries, descriptions of intended uses, and results from evaluations and testing - see <https://docs.cohere.com>

15.B. If yes, how often are such reports usually updated?

At or before major product releases

16. How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?

- Public documentation
- Protected information sharing
- Regulatory reporting as relevant
- Direct engagement with AI safety institutes, civil society or academic partners

Describe how evaluation results are shared and with which stakeholders (e.g. other organizations, governments, civil society and academia, etc). You may also indicate whether non-technical or plain-language summaries are provided for broader audiences.

- Some evaluation results are shared publicly - e.g. when new models are announced (see for example - <https://cohere.com/blog/command-a>).
- Some evaluation results are shared privately with customers and partners, especially where we are developing an AI model or system with or for them, for example via fine-tuning.
- We share information with other parties, such as civil society or government actors on request and through informal engagements.

17. Does your organization publish public privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

Yes

If yes, provide links to privacy policies or notices covering personal data, user inputs, and AI-generated outputs.

All privacy policies and policies relevant to data, inputs, and outputs are available at:
<https://cohere.com/model-training-privacy-notice>

18. How does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

- Public documentation
- Provide information or tools to downstream actors on data quality
- Protected information sharing
- Regulatory reporting as relevant

Please provide links to any publicly available information.

The Public Summary of Training Content, per EU AI Act Article 53(1)(d) is available at
<https://docs.cohere.com/docs/command-a-plus>.

19. Does your organization demonstrate transparency related to advanced AI systems through any other methods?

Yes

If yes, describe additional method(s), including where relevant transparency measures related to AI agents

Cohere publishes comprehensive information and documentation about our AI models, via model cards, technical reports, and the documentation section of our website; provides educational resources like Cohere's LLMU to support AI literacy; releases many of our models with open weights and permissible licensing to support innovation, as well as third-party research and inspection.

20. Does your organization publish statistical insights derived from anonymised, aggregated usage data to advance understanding of AI adoption and its economic impacts in areas including occupations and tasks by region, sector and country? To this end, do you contribute to international cooperation efforts?

No

Please estimate the total staff time and annual expenses associated with conducting compliance activities and include any further comments and implementation documentation (optional).

No answer provided

Section 4 – Organizational Governance, Incident Management, and Transparency

21. How has AI risk management been embedded in your organization's governance framework?

- Board of Directors oversight
- Executive oversight
- Senior management oversight
- Governance council or committee

Please describe how AI risk management is integrated into governance structures, including: which roles or bodies have advisory versus decision-making authority; which roles or bodies have final accountability for AI risk-related decisions; and how often policies are reviewed or updated (e.g., following major incidents, regulatory changes, or material updates to systems).

Risk management is integrated into governance at Cohere via:

- A cross-functional leadership committee that provides internal oversight and strategic direction for Cohere's Artificial Intelligence Management System (AIMS), ensuring executive support and review of Cohere's AI governance practices.
- A cross-functional team of technical and non-technical staff from the Policy, Legal, Safety, Data and Modeling, Research, and Security teams. This group is responsible for operationalizing governance

mechanisms in their respective functions, such as developing policies, conducting security audits, carrying out safety testing, sharing insights and developments in AI governance, and advancing AI governance objectives at Cohere.

- Clear, final points of evaluation to validate that a model is safe and secure, and that accountability for decisions to make a product available to our customers is escalated to our C-suite leaders.

For more information, see: <https://cohere.com/blog/building-trust-in-ai-coheres-approach-to-ai-governance>

22. Are relevant staff trained on your organization's governance policies and risk management practices?

Yes

If yes, describe training methods (e.g., onboarding modules, regular sessions, and confidential channels to report concerns).

- All staff are provided with initial and ongoing security training and are required to annually review and acknowledge policies and procedures, including with respect to AI and risk management.
- Cohere staff are provided with access to our Cohere Academy, an internal certification that provides guidance, resources, and exercises to staff to transform how they work with AI, by leveraging Cohere's 'North' Chat, Agents, and Automations. The certification tests both technical North proficiency and the essential mental models needed to truly adopt AI as a routine part of your work.
- Our safety, security and legal teams have open channels on our internal communications platform that any employee can use to find information or report concerns, in addition to formal whistleblower channels made available to all staff.

23. Are steps taken to address reported incidents documented and maintained internally?

Yes

If yes, describe how incident records are kept and used for follow-up (e.g., incident logs, root-cause analysis reports, internal tracking systems). You may indicate whether and how you define escalation thresholds for when incidents must be surfaced to senior leadership or external parties.

Cohere maintains a documented Incident Response Policy with a dedicated cross-functional Incident Response Team (IRT) that any employee can initiate by notifying the internal security team through multiple channels. The structured process includes identifying, prioritizing, communicating, assigning, and tracking incidents through resolution, followed by formal "Lessons Learned" documentation and periodic testing via tabletop exercises to continuously improve the response capabilities.

24. Does your organization share research and best practices on addressing or managing risk?

Yes

If yes, describe how research or good practices are shared (e.g., publications, conferences, collaborations with other sectors).

Cohere has published papers and information via our website about our own approach to risk management and our approach to AI Governance. We have also published several research papers and policy papers relating to AI safety. We share our expertise informally with partners, and at professional, business, and technical conferences.

25. Does your organization use international technical standards or best practices for AI risk management and governance policies?

Yes

If yes, please indicate which standards or frameworks apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001)
- Industry-led best practices or codes of conduct

If yes, describe which international technical standards or best practices your organization uses.

Cohere's AI governance approach is aligned with and certified under ISO/IEC 42001, which ensures systematic and ongoing risk management of AI systems, robust governance practices that are regularly reviewed, and external validation steps.

Our approach is also driven by international voluntary standards for AI developers, which provide principles-based frameworks that foster harmonization and trust across the AI industry. Since 2023, Cohere has signed onto several voluntary commitments and frameworks for AI governance, including:

- The Seoul Frontier AI Safety Commitments and the Seoul AI Business Pledge.
- Canada's Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems
- The EU AI Act's voluntary General-Purpose AI Code of Practice for general-purpose AI models, the Transparency of AI Generated Content, and the AI Pact

We also contribute to and align our approach with other leading frameworks for AI governance.

Any further comments and implementation documentation.

No answer provided

Section 5 – Content Authentication & Provenance Mechanisms

26. What mechanisms, if any, does your organization put in place to allow end-users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?

- User disclosure (e.g., text notice, verbal statement, or system label)

- Documentation or FAQs explaining when and how AI is used in the system

Provide any additional details or context on how or when users are informed they are interacting with advanced AI systems (e.g. if such mechanisms are planned, in testing, or applied only to certain products).

Cohere is a signatory to the EU AI Act Code of Practice for the Transparency of AI Generated Content, and will implement additional marking measures in accordance with the requirements of the code.

27. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems?

Yes

27.A. If yes, how?

- Latent disclosure (e.g., hidden provenance information)
- Manifest disclosure (e.g., visible label)
- Use of international or industry standards (e.g., C2PA)
- Other (specify): Cohere is a signatory to the EU AI Act Code of Practice for the Transparency of AI Generated Content, and will implement additional marking measures in accordance with the requirements of the code and per the timelines set out in the AI Act.

27.B. If yes, does your organization use international technical standards or best practices when developing or implementing content provenance?

Yes

27.C. Does your organization collaborate on or invest in research to advance content authentication and provenance?

Yes

Describe how provenance mechanisms are implemented and which standards or tools are used, and any collaborations or investments to improve content authentication (e.g., C2PA credentials, joint R&D projects, open-source initiatives, standard-setting groups). You may also indicate whether these mechanisms are tested for robustness against removal, tampering, or circumvention.

Cohere is currently advancing internal research for provenance mechanisms that work across text outputs and documents, drawing on industry best practices and open-source tools.

Any further comments and implementation documentation.

No answer provided

Section 6 – Research & Investment to Advance AI Safety & Mitigate Societal Risks

28. Does your organization advance research and investment in the following areas?

- Security
- Safety
- Data quality and representativeness
- System performance and outcomes (e.g. accuracy, consistency)
- Explainability or interpretability
- Robustness or reliability
- Other (please specify): Multilingual and multicultural AI model performance and safety; AI and the future of work.

If yes, please describe key research or investment initiatives in the areas selected above (e.g., dedicated safety teams, funding of research programs, contributions to open-source safety tools, publishing research, partnerships and collaborative research initiatives). You may also indicate whether related outputs are publicly shared, shared with partners or within consortia or not publicly disclosed.

Cohere's open science research team - Cohere Labs - have published over 100 papers as well as several models and datasets released under research licenses since being established in 2024. Much of this work has focused directly or indirectly on safety, and has been produced through an annual scholars program and multiple community science projects involving thousands of contributors from around the world, to expand who has access to opportunities to advance machine learning research. Additionally, members of our Safety and Security teams have collaborated on research papers, presented at conferences, and participated in working groups and collaborative efforts to advance AI science and safety.

29. Does your organization pursue research or investment related to socio-economic or environmental considerations in the development and deployment of AI?

- Socio-economic considerations (e.g., AI adoption and productivity, workforce transitions, skills development, access to AI tools)
- Environmental considerations (e.g., energy use, computational efficiency, resource use)

If yes, please provide examples of relevant research or investment initiatives for the areas selected above.

Our 'Future(s) of Work' program focuses on the impact of AI on work and jobs:

<https://cohere.com/research/futures-of-work>

Cohere's models are designed to be computationally efficient, ensuring the scale of the model matches that of the task, and we have published several studies and papers on computational efficiency:

<https://cohere.com/research/papers>

Any further comments and implementation documentation.

No answer provided

Section 7 – Advancing Human and Global Interests

30. Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities and limitations of advanced AI systems?

Yes

If yes, describe any digital literacy or training initiatives your organization supports (e.g., user education, public AI-education programs, school or university partnerships, community workshops). You may also indicate whether these efforts include or prioritise underserved or underrepresented communities.

Examples of Cohere's efforts to support AI literacy and training include:

- LLM University - Cohere's premier learning destination for mastering Enterprise AI technologies. Designed for developers and technical professionals, this hub offers comprehensive resources, expert-led courses, and step-by-step guides to help people start building quickly and stay ahead in the rapidly evolving AI landscape. <https://cohere.com/llmu>
- Partnerships with several academic institutions to support research and learning, for example with the University of Waterloo to strengthen Canada's AI talent pipeline, by preparing students to help organizations put AI to work on real business challenges. <https://cohere.com/blog/cohere-university-of-waterloo-announcement>

31. Does your organization pursue AI projects that aim to deliver societal or public benefits, including through partnerships with civil society and community groups?

Yes

If yes, please provide examples, including any collaborations with civil society, community groups, or efforts to address broader societal challenges.

Cohere Labs' Catalyst Grants enable academic partners, civic and public impact organizations to access Cohere's models and apply them to social-benefit projects. We have awarded more than 250 grants across 35 countries, totalling \$350k of API credits. <https://cohere.com/research/grants>

Any further comments and implementation documentation.

No answer provided