# OECD AI Transparency Report

**Organization:** Salesforce (USA) **Reporting Period:** Q2 2025 **Published:** April 15, 2025

## Section 1 – Risk identification and evaluation

**a. How does your organization define and/or classify different types of risks related to AI, such as unreasonable risks?**

In general, in the enterprise space, where customers can tailor and use AI systems in various contexts, the most effective approach is for developers to perform a comprehensive assessment of reasonably foreseeable potential adverse impacts from the use of the system.

Embedding trust into Salesforce's models, products, and features requires close collaboration with our Technology and Product teams. At the heart of this effort is the Trusted AI Review process, led by Responsible AI and Tech (RAIT) product managers within Salesforce's Office of Ethical and Humane Use (OEHU). This process enables potential risks to be identified early, mitigated effectively, and tracked transparently.

During the review, RAIT product managers work closely with product teams to understand the product's use cases, tech stack, and intended audience. They conduct a risk assessment to identify and categorize potential risk scenarios under sociotechnical harm subtypes. For each risk vector, the inherent and residual risk are identified so as to have a complete and clear understanding of its impact.

Our AI Acceptable Use Policy (AI AUP) includes a detailed list of uses for which our customers are not allowed to use our generative AI tools, including:

- Automated decision-making processes with legal effects
- Individualized advice from licensed professionals
- Explicitly predicting an individual's protected characteristics
- Engaging in coordinated inauthentic behavior for the purposes of manipulating public debate & political campaigns
- Child sexual exploitation and misuse
- Weapons development
- Adult content and dating apps

Additionally, our AI AUP specifies that our generative AI tools are not intended for high risk uses that could result in the death or serious bodily injury of any person or in other catastrophic damage, including through warfare or the operation of critical infrastructure.

**b. What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks and misuse, throughout the AI lifecycle?**

We evaluate our AI models against trust and safety metrics (e.g., bias, privacy, robustness) to ensure that they perform at the highest level. When a model scores below a certain range on one of these metrics, we use adversarial testing to better understand how that manifests in practice. For example, knowing the exact toxicity score does not reveal what kind of toxic content a LLM might generate. However, once you identify toxicity as a potential issue, you can focus your adversarial testing specifically on that, rather than testing for all types of risks.

Our Responsible AI & Technology team implements red teaming practices to improve the safety of our AI products. The team has conducted several end-to-end adversarial tests to prevent bias, toxicity, and ensure alignment to ethical tech commitments for our own models and applications as well as those of our partners.

Our AI services are also part of our vulnerability and incident management processes under security. Salesforce's Product Security Strategy & Advisory Security team proactively identifies and evaluates AI risks throughout the lifecycle using design reviews and penetration testing across their environments and specifically for Agentforce. Their "ExploitAI" activities are crucial for uncovering and validating AI-specific vulnerabilities beyond traditional security flaws. This includes focusing on authorization, system interaction, manipulation of goals and instructions, hallucination exploitation, impact analysis, knowledge poisoning, memory and context issues, multi-agent exploitation, resource exhaustion, and supply chain attacks. Recognizing that AI systems exhibit behavioral changes rather than simple failures, their testing evaluates a spectrum of responses, explicitly addressing concerns like bias, "jailbreaks," and data access violations. This comprehensive approach combines standard security practices with AI-focused testing to ensure the confidentiality, integrity, and availability of our AI features.

**c. Describe how your organization conducts testing (e.g., red-teaming) to evaluate the model's/system's fitness for moving beyond the development stage?**

There are two main ways to go about red teaming — manual and automated — both of which are employed at Salesforce.

Manual testing leverages human testers who think like adversaries, using their expertise to craft complex and sophisticated attack strategies that automated systems might overlook. Examples include hackathons (for example, Salesforce's XGen Hackathon had teams compete to identify vulnerabilities in our next-generation text generation models); or bug bounties (these are excellent once a product is launched to catch new harms that weren't discovered during pre-launch. We incentivize our employees to identify and report vulnerabilities through our Bug

Bounty Program and host ethical bug bounties. To date, Salesforce has invested more than $23M in our bug bounty program.)

As of September 2024, our teams have conducted 19 internal and 2 external red teaming exercises across our suite of generative AI models and applications. Through pre-launch testing, we have reduced toxic, biased, and/or unsafe outputs by 35% in an AI marketing feature; added guardrails for AI Agents to prevent bias and increase transparency; and partnered with one of our systems integrator partners to decrease biased outputs in generated content from 69% to 4%.

Automated testing is used as an enhancement, not replacement, of human-driven testing and evaluation. This type of testing involves the use of scripts, algorithms, and software tools to simulate a vast number of attacks or adversarial scenarios in a short period, systematically exploring the risk surface of the system. One approach we've been taking to automate some of our tests is called "fuzzing," where we generate randomized test cases based on successful human attacks from manual testing (confirmed by to have been successful either in our manual testing, or through other publicly known attacks), deliver these test cases to the target model and collects outputs, and then assess whether each test case passed or failed.

Another way we test our products is by performing precautionary "prompt injection attacks", by crafting prompts specifically designed to make an AI model ignore previously established instructions or boundaries. Anticipating actual cybersecurity threats like these is essential to refining the model to resist actual attacks.

Employee Trust Testing: To surface more subtle forms of bias that users experience, we tapped employees from across Salesforce's global workforce to evaluate the trustworthiness of prompt responses. Participants engaged with the AI in simulated real-world scenarios, creating detailed personas to represent varied user experiences and using those personas to explore interactions that probed for biases and inconsistencies.

Salesforce has developed the world's first LLM benchmark for CRM to assess the efficacy of generative AI models for business applications. This benchmark evaluates LLMs for sales and service use cases across accuracy, cost, speed, and trust and safety based on real CRM data and expert evaluations. What sets this benchmark apart is the human evaluations by both Salesforce employees and actual external customers and the fact that it is based on real-world datasets from both Salesforce and customer operations.

**d. Does your organization use incident reports, including reports shared by other organizations, to help identify risks?**

Yes

**e. Are quantitative and/or qualitative risk evaluation metrics used and if yes, with what caveats? Does your organization make vulnerability and incident reporting mechanisms accessible to a diverse set of stakeholders? Does your organization have incentive programs for the responsible disclosure of risks, incidents and vulnerabilities?**

Trust and safety (T&S) benchmarking is a critical aspect of our evaluation process. By evaluating our AI models against T&S metrics (e.g., bias, privacy, robustness), we can ensure that they perform at the highest level. When a model scores below a certain range on one of these metrics, we use adversarial testing to better understand how that manifests in practice.

We also use the quantitative T&S metrics in this paper "TrustLLM: Trustworthiness in Large Language Models".

In terms of caveats, we would like to note that all quantitative metrics are limited in communicating the nature of a particular risk. For example, a specific toxicity score cannot communicate how often a model or system might generate toxic content or the severity of that toxicity. Additionally, some models may have been trained on publicly available evaluation datasets and, therefore, may score especially well on a benchmark (i.e., overfitting). Adversarial testing can provide qualitative data to help organizations understand exactly what kind of toxicity is generated and how easily it is generated (e.g., with expected use or only via significant attacks).

**f. Is external independent expertise leveraged for the identification, assessment, and evaluation of risks and if yes, how? Does your organization have mechanisms to receive reports of risks, incidents or vulnerabilities by third parties?**

Regarding external independent expertise:

- We have engaged experts to perform penetration tests (through our Security Team's Bug Bounty program). We also recently chose to outsource testing of two of our Einstein for Developers (E4D) products and our research multimodal model, PixelPlayground.
- The Ethical Use Advisory Council is our overarching body that guides the Office of Ethical and Humane Use in its product and policy recommendations to leadership. This Advisory Council was established in 2018 and is composed of external experts from academia and civil society along with internal VP+ level executives and frontline employees.
- Through our Responsible Disclosure Policy, we encourage responsible reporting of any vulnerabilities that may be found in our site or applications. Salesforce remains committed to working with security researchers to verify and address any reported potential vulnerabilities.

Regarding mechanisms to receive reports of risks, incidents or vulnerabilities by third parties:

- Incidents and vulnerabilities can be discovered and reported by our customers with the help of the Einstein Trust Layer, which includes audit trails for use in third-party reporting.

Customers get a full audit trail of their generative AI transactions (prompts, responses, trust signals, user feedback) in their Data Cloud, fully under the customers' control, for them to use in their monitoring.

- We also have feedback mechanisms within the AI features so customers can indicate whether the generated output was inaccurate, inappropriate, etc.
- Customers, partners, and the general public can submit reports concerning incidents and vulnerabilities through our security email alias, as well as through the bug bounty program.

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

Salesforce employees participate in several working groups of the NIST Artificial Intelligence Safety Institute Consortium (AISIC). The Consortium brings together more than 280 organizations to develop science-based and empirically backed guidelines and standards for AI measurement and policy.

A full list of Salesforce's compliance certifications and attestations can be found here.

We published Sustainable AI Policy Principles, which recommend the consideration of environmental impact when determining the risk of AI systems and the establishment of energy efficiency standards for high-risk systems. We also co-developed the AI Energy Score project which introduces such a standard.

**h. How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks?**

At Salesforce, we are committed to knowledge sharing across industry, government, and civil society to advance trusted AI in society, and regularly share our principles, practices, and learnings.

In 2024, we published 20+ blog posts dedicated to the ethical and humane use of AI ranging from the top risks and related guidelines for generative AI to how we've built trust into our AI.

To help our customers dive deeper, we've also created resources and guides like the National Institute of Standards and Technology (NIST) AI Risk Management Framework quick-start guide and our Human at the Helm action pack. We also published the world's first LLM Benchmark for CRM, which includes trust and safety metrics for each model.

Members of our Office of Ethical and Humane Use have held active positions on various AI councils around the globe, including the U.S. National AI Advisory Committee; the U.S. Chamber

of Commerce Commission on Artificial Intelligence Competitiveness, Inclusion, and Innovation; Singapore's Ethical AI Advisory Council; the Freedom Online Coalition's Advisory Network; the Washington State Artificial Intelligence Task Force; the Oregon State Taskforce on Artificial Intelligence; and the U.S. National Institute of Standards and Technology (NIST). We are also an active member of the UNESCO Global Business Council for the Ethics of AI and Singapore's AI Verify AI Foundation.

We are proud to have endorsed several global voluntary commitments. These include pledges to conduct internal and external testing before product release, sharing information on risks and vulnerabilities with each other and government entities, and conducting research to address societal risks. To that end, we have signed onto the following voluntary commitments and industry pledges:

- The EU AI Pact;
- Canada's Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems;
- The Seoul AI Business Pledge;
- The Trento AI Declaration.

Salesforce is also an active member of several industry alliances, including the Data & Trust Alliance, Data Provenance Initiative, AI Alliance, WEF AI Governance Alliance, Sustainable AI Coalition, and the WEF Global Future Council on Data Equity. By engaging in these partnerships, we collaborate with other technology leaders to develop standards, frameworks, and practices that promote responsible AI use across sectors, addressing data integrity, bias mitigation, and sustainable AI deployment.

**Any further comments and for implementation documentation**

We are grateful to the G7 and the OECD for their work on the reporting framework.

Salesforce is the #1 AI CRM, helping companies connect with their customers in a whole new way. We pioneered cloud-based CRM in 1999, and today we are leading the shift to trusted, agentic AI. With Agentforce, Salesforce enables organizations to deploy autonomous AI agents that act on unified, real-time data across their systems, helping every employee deliver more personalized, efficient, and secure customer experiences. Our trusted platform powers AI, data, and CRM applications across sales, service, marketing, commerce, and IT, so every team can work smarter and drive meaningful business outcomes.

Salesforce's commitment and innovative leadership on responsible and trustworthy AI is inspired by the high standards required for successful outcomes by our enterprise customers around the world. Our enterprise customers require AI that performs at the highest levels of trust and safety, and that addresses their priorities – accuracy, robustness, auditability, privacy, security, and toxicity and bias mitigation.

## Section 2 – Risk management and information security

**a. What steps does your organization take to address risks and vulnerabilities across the AI lifecycle?**

Our Responsible AI & Technology team works with product teams and technical experts to develop and assess mitigations for identified risks. When additional testing is needed, they collaborate with the Testing, Evaluation, and Alignment team to validate mitigations. This process leverages both qualitative judgment and defined thresholds to ensure consistent, repeatable assessments. All review details are documented in a standardized Trusted AI assessment template and tracked using Salesforce's internal tools. Examples are included in our Trusted AI and Agents Impact Report.

Salesforce has standardized the guardrails implemented across our AI products, designed to improve safety, accuracy, and trust while empowering human users. We call these trust patterns. Examples include:

- Mindful Friction: Mindful friction is a design pattern in responsible AI that intentionally incorporates checkpoints or pauses within AI workflows to encourage thoughtful decision-making and reduce the risk of unintended consequences. Unlike traditional automation, which prioritizes speed and efficiency, mindful friction encourages critical actions involving AI to be reviewed by a human who may take additional validation steps.
- Transparency and Notification: Users are informed whenever they interact with an AI system, comprehend the role of AI in generating outputs, and are aware of its capabilities and limitations. Notifications and disclosures, such as those embedded in Salesforce's AI agents, clarify whether communications or actions are AI-generated.

The Einstein Trust Layer is a comprehensive framework within Salesforce's AI ecosystem, designed to uphold data privacy, security, and ethical standards while enhancing the effectiveness of AI applications. Its core functionalities include:

- Secure data handling
- Zero data retention
- Ethics by design
- Audit Trail
- Real-Time Toxicity Detection (more details on all these functionalities are included in our Trusted AI and Agents Impact Report).

We make publicly available our Acceptable Use Policy and AI Acceptable Use Policy, and other public-facing materials as outlined above.

Our security team has published best practices for how organizations can protect themselves from AI security risks and vulnerabilities, such as prompt injections, data poisoning, and supply chain vulnerabilities.

**b. How do testing measures inform actions to address identified risks?**

We feed the results of our benchmarking or adversarial testing back to the relevant teams to either improve the model or to create new guardrails.

**c. When does testing take place in secure environments, if at all, and if it does, how?**

All of the testing we do is within Salesforce, using non-confidential data, and before the product launches.

**d. How does your organization promote data quality and mitigate risks of harmful bias, including in training and data collection processes?**

Our Office of Ethical and Humane use has stood up an ethical testing and assurance team to test our models and features for biased and toxic outputs. Their work consists of both probing for edge cases (malicious use) as well as testing for undesirable outputs when used as expected (benign misuse).

We evaluate our models and apps against a sociotechnical harms framework to consistently score risks based on severity and frequency.

We conduct multiple types of benchmarks and adversarial tests to evaluate our models on toxicity, bias, privacy (data leakage), robustness, and ethics.

We feed the output of our adversarial testing into our models to conduct “unlearning,” a form of reinforcement learning with human feedback (RLHF).

Our AI Research team works closely with our Office of Ethical and Humane Use to curate new datasets for the Einstein Trust Layer detectors for toxicity and bias.

Our legal team has been diligent about not allowing models going into products to be trained on known harmful datasets. We don't scrape the web, and certain datasets are banned.

We have also published several blog posts talking about the risks of AI, including our guidelines for the responsible development of generative AI. We also have a number of modules on Trailhead, our free online learning platform, on responsible AI and generative AI.

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

We use legally compliant datasets to train our models to respect data provenance.

Both our general Acceptable Use Policy and our AI Acceptable Use Policy include specific provisions to protect third parties' intellectual property rights.

**f. How does your organization protect privacy? How does your organization guard against systems divulging confidential or sensitive data?**

Trust is our number one value at Salesforce and our customers' data is not our product. Our top priority is the security and privacy of the data that we are entrusted to protect. Our AI systems are built and trained on curated data with strict requirements for privacy, security, and accuracy, and we make available a number of relevant safeguards, which can include:

- Dynamic grounding steers a LLM's answers using the correct and the most up-to-date information, "grounding" the model in factual data and relevant context. This prevents AI hallucinations, or incorrect responses not grounded in fact or reality.
- Toxicity detection is a method of flagging toxic content such as hate speech and negative stereotypes. It does this by using a machine learning model to scan and score the answers a LLM provides, ensuring that generations from a model are usable in a business context.
- Zero retention means that no customer data is stored outside of Salesforce. Generative AI prompts and outputs are never stored in the LLM, and are not learned by the LLM.
- Auditing tools allow organizations to evaluate systems to make sure they are working as expected, without bias, with high quality data, and in-line with regulatory and organizational frameworks. Auditing also helps organizations meet compliance needs by logging the prompts, data used, outputs, and end user modifications in a secure audit trail.
- Retrieval-augmented generation (RAG) is a technique that uses semantic search to retrieve relevant snippets of information from any data source for better AI outputs. In accordance with global privacy principles, companies should only use the minimum amount of personal data needed for a task. RAG and similar techniques can automate and scale the process of determining the minimum set of relevant data for grounding, giving organizations the benefit of higher-quality outputs that rely on just the right amount of external data.

We have published more information about how agentic AI systems like Agentforce can enhance privacy protections here.

**g. How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security?<br><ul><li>i. How does your organization assess cybersecurity risks and implement policies to enhance the**

**cybersecurity of advanced AI systems?</li><li>ii. How does your organization protect against security risks the most valuable IP and trade secrets, for example by limiting access to proprietary and unreleased model weights? What measures are in place to ensure the storage of and work with model weights, algorithms, servers, datasets, or other relevant elements are managed in an appropriately secure environment, with limited access controls in place?</li><li>iii. 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?</li><li>iv. How often are security measures reviewed?</li><li>v. Does your organization have an insider threat detection program?</li></ul>**

How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security?

- The security of our customer data and our products is our top priority, and we have specific measures in place that ensure robust operational, cyber, and physical security.
- For higher risk generative AI features and use cases, we have a robust review process that includes approvals from security, ethics, and product legal teams, among others.
- When it comes to open-source models, we first run them on our own equipment or cloud accounts to ensure they comply with licensing guidelines.
- For features that allow customers to use their own model, we provide secure storage for API keys and transparently disclose when customers are interacting with automated systems.
- By default, we ensure human oversight is still required for high-risk or high-judgment decisions, and we apply strong security measures to protect system access.
- We also maintain distinct trust zones and ensure that access keys are not shared between production and non-production systems.
- We have designed our incident response plans to address a wide range of risks, especially in the age of AI.
- We build our products with a security-first mindset, including threat modeling that treats AI-generated content as crossing a trust boundary.
- We provide each Salesforce customer with the default security, out-of-the-box tools and educational resources necessary to protect their data. That’s why we have our Einstein Trust Layer with zero-data retention safeguards built into the platform.
- We also serve as a trusted advisor to our customers to help ensure they are also following security best practices and adopting proper controls, such as multi-factor authentication (MFA) and the principle of least privilege.

How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems?

- Our Security organization includes a Product Security Strategy & Advisory team, charged with ensuring that the existing and emerging risks that pose the greatest threats to our business and customers are taken into account throughout the company. The team performs in-depth discovery work aligned to the phases of the delivery lifecycle (secure design, development,

and run) and red teaming to protect against key security risks like data exfiltration and prompt injection. We also maintain an insider threat program for our general operations.
- The Open Web Application Security Project (OWASP) recently revealed the top 10 LLM risks: prompt injections, insecure output handling, training data poisoning, model denial of service, supply chain vulnerabilities, sensitive information disclosure, insecure plugin design, excessive agency, overreliance, model theft. You can find a detailed account of how Salesforce is addressing the security risks here.

How does your organization protect against security risks the most valuable IP and trade secrets, for example by limiting access to proprietary and unreleased model weights? What measures are in place to ensure the storage of and work with model weights, algorithms, servers, datasets, or other relevant elements are managed in an appropriately secure environment, with limited access controls in place?

- Access to Salesforce-hosted models is implemented in accordance with all other Salesforce production elements and conforms to existing security standards and best practices. The controls used include requiring Just In Time (JIT) credentials, Multi-factor Authentication (MFA), strong audit trails, and logging.
- Requests to the models used by Einstein are required to pass through the Einstein Trust Layer, which implements strong authentication and authorization controls through OAuth. This ensures that only authenticated and authorized clients can access the model. We also use prompt defenses, rate-limiting, and watermarking to prevent theft of intellectual property and detect content for monitoring and alerting purposes.
- Our AI, Data, and CRM products are hosted on Salesforce's infrastructure platform, Hyperforce. Products running on Hyperforce benefit from its integration of enhanced standards for compliance, security, agility, and scalability, and from Salesforce's continued commitment to privacy.

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?

- Salesforce has a rigorous vulnerability management program to identify, assess, and remediate vulnerabilities across its products and services, assisting in the protection of customer data and data we hold as a controller. The program includes internal and third-party assessments, a bug bounty program, continuous monitoring, and engaging in threat intelligence information sharing programs to help surface and remediate known issues in third party hardware and software as early as possible in the vulnerability lifecycle. Salesforce uses a risk-based approach to prioritize and remediate identified vulnerabilities, following industry standards and best practices.

How often are security measures reviewed?

- Security measures in our organization are reviewed at various intervals to ensure they remain effective and up-to-date. Information security policies, standards, and controls are

reviewed annually. Our Information Security Management System (ISMS) Review Meetings are held semi-annually. Production Access (including access to our AI services) reviews are conducted quarterly, while privileged accounts are reviewed monthly. Separation of duties and privileges are reviewed every 30 days. Additionally, our security controls undergo continuous monitoring and are audited by external auditors for various security and compliance certifications at least annually.

Does your organization have an insider threat detection program?

- Salesforce has an insider threat program to monitor for potential risks posed by employees, contractors, or other insiders with privileged access to our systems. Salesforce also has a threat detection program designed to help our customers identify potential insider threats with access to their Salesforce data.

**h. How does your organization address vulnerabilities, incidents, emerging risks?**

Our Security, AI Research, and Ethical and Humane Use teams have conducted focused bug bounties for product features and models, including, for instance, a recent bug bounty on our Einstein for Developers model, which you can read more about here.

In addition to bug bounties, issues can be surfaced and reported with the help of the Einstein Trust Layer, where users can report vulnerabilities through the audit trail.

Internally, the Salesforce Incident Response Team actively monitors an extensive array of public and private threat intelligence feeds and communities. This intelligence is processed in near-real time, enabling the team to swiftly detect and respond to emerging threats.

Beyond internal monitoring and bug-bounty programs, Salesforce encourages external parties - including vendors, individual researchers, and customers - to report identified vulnerabilities. This can be done by following the guidelines outlined in the Security Vulnerability Finding Submittal Guide. Reported vulnerabilities are then assessed and assigned an internal ranking based on the OWASP risk rating framework, which considers both the likelihood and impact of the threat. These vulnerabilities are meticulously tracked through to resolution, adhering to company policies and industry best practices.

**Any further comments and for implementation documentation**

*No answer provided*

## Section 3 – Transparency reporting on advanced AI systems

**a. Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?<br><ul><li>i. How often are such reports usually updated?</li><li>ii. How are new significant releases reflected in such reports?</li><li>iii. Which of the following information is included in your organization's publicly available documentation: details and results of the evaluations conducted for potential safety, security, and societal risks including risks to the enjoyment of human rights; assessments of the model's or system's effects and risks to safety and society (such as those related to harmful bias, discrimination, threats to protection of privacy or personal data, fairness); results of red-teaming or other testing conducted to evaluate the model's/system's fitness for moving beyond the development stage; capacities of a model/system and significant limitations in performance with implications for appropriate use domains; other technical documentation and instructions for use if relevant.</li></ul>**

Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?

- Salesforce’s publications in this area include the following:
- Public documentation on product risks and mitigations on the Salesforce Newsroom/blogs, our free online learning platform Trailhead, product help documentation (examples: Understand Salesforce Use Cases and Limitations, Einstein Generative AI Features), in-app warnings/models, and external publications.
- Our Trust & Compliance documentation.
- Our Trusted AI and Agents Impact report.
- Our Acceptable Use Policy and AI Acceptable Use Policy.
- Model cards for several of our predictive AI models to disclose the risks of the technology, ethical considerations, use cases, and more.
- Our Human at the Helm resources.

How often are such reports usually updated?

- Generally, on a regular basis.
- Updates to our Trust & Compliance documentation are reflected in this Change Log.

How are new significant releases reflected in such reports?

- As an example, all relevant documentation has been updated to reflect our latest advancements in agentic AI with Agentforce.

Which of the following information is included in your organization’s publicly available documentation: other technical documentation and instructions for use if relevant.

**b. 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?**

We publish relevant information on our website, in particular with regard to our red teaming practices. Examples can be found here, here, here, and here.

**c. Does your organization disclose privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?**

Salesforce gives customers control over the use of their data for AI. Whether using our own Salesforce-hosted models or external models that are part of our Shared Trust Boundary, no context is stored. The large language model forgets both the prompt and the output as soon as the output is processed.

Our Einstein Trust Layer includes seamless privacy and data controls, including secure data retrieval, zero data retention, and dynamic grounding.

The Salesforce Privacy Information page contains information around how Salesforce protects Personal Data of Customers and as a Controller.

**d. 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?**

In line with our guiding principle of transparency, we work to ensure that our models and features respect data provenance.

As outlined in our white paper “Mitigating LLM Risks Across Salesforce’s Gen AI Frontiers”, Salesforce’s research teams develop AI models using open-source datasets that are carefully vetted by Salesforce to meet Salesforce standards, legal obligations, and ethical objectives. Additionally, researchers can use anonymized and aggregated data in training these models, which is only done with explicit approval and agreement from customers whose data is included. Salesforce developers comply with industry standards including but not limited to OWASP 10, CWE25, NIST AI RMF.

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

Transparency and honesty are core tenets of our trusted AI principles and our guidelines for trusted generative AI.

Under Section 2 (a) of the reporting framework, we described our trust patterns, including on transparency and notification, which help AI systems operate within predefined boundaries and maintain transparency.

Salesforce's Audit Trail is designed to provide transparency and accountability in AI operations. It allows admin users to track the actions and decisions of AI systems. By documenting the entire lifecycle of an AI interaction, the Audit Trail provides a detailed log of what actions the AI agent performed, why those actions were taken, and what data or inputs influenced those decisions.

Our LLM Benchmark for CRM, the first of its kind globally, gives organizations a sense of transparency regarding trust and safety.

Our participation in international commitments and business pledges, including Canada's Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems, the Seoul AI Business Pledge, and the Trento AI Declaration, reflect our pledge to uphold high ethical standards in AI, aligning with global efforts to foster transparency and accountability.

Salesforce discloses environmental impacts of internally developed AI models, both pre-training impacts and inference energy use via the AI Energy Score project (co-developed by Salesforce).

**Any further comments and for implementation documentation**

*No answer provided*

## Section 4 – Organizational governance, incident management and transparency

**a. How has AI risk management been embedded in your organization governance framework? When and under what circumstances are policies updated?**

We developed our first set of trusted AI principles in 2018, and we continue to guide the responsible development and deployment of AI at Salesforce through principles, policies, and products.

As we entered into the era of generative AI in early 2023, we augmented our trusted AI principles with a set of 5 guiding principles for developing responsible generative AI as the first enterprise company to put out guidelines in this emerging space.

As AI continues to evolve, and we enter the space of agentic AI, these principles still hold true. We are focused on intentional design and system-level controls that enable humans and AI to work together successfully and responsibly.

In addition to principles and policies, having the right decision-making structures in place is critical to ensuring the responsible development and deployment of AI. At Salesforce, there are many governance and accountability structures for trusted AI. To note a few:

- Salesforce's Cybersecurity and Privacy Committee of the Board of Directors meets with our Chief Ethical and Humane Use Officer quarterly to receive updates and provide feedback on key trusted AI priorities.
- The Office of Ethical and Humane Use also has regular interactions with the executive leadership team of the company to discuss policy and product topics for review and approval.
- The Human Rights Steering Committee meets quarterly. It includes executives from the Office of Ethical and Humane Use, Legal, Privacy, Employee Success, Procurement, Government Affairs, Equality and Sustainability, who jointly oversee our human rights program, including efforts to monitor, identify and mitigate salient human rights risks.
- The AI Trust Council comprises executives across Security, Product, Engineering, AI Research, Product Marketing, Legal, UX, and Ethical and Humane Use. The Council was formed to align and speed up decision-making for AI products and meets on a bi-weekly basis.
- The Ethical Use Advisory Council is our overarching body that guides the Office of Ethical and Humane Use in its product and policy recommendations to leadership. This Advisory Council was established in 2018 and is composed of external experts from academia and civil society along with internal VP+ level executives and frontline employees (below VP level). The Advisory Council meets quarterly and provides strategic guidance, feedback, and counsel on the top priorities of the Office of Ethical Use.

**b. Are relevant staff trained on your organization's governance policies and risk management practices? If so, how?**

Relevant information, training sessions and material on responsible AI & tech are distributed to relevant employees through our internal Slack channels maintained by our Office of Ethical and Humane Use.

All employees undergo annual security training, including training on how to maintain data integrity and confidentiality. Salesforce conducts role-based training for engineers, developers, business technology and roles with access to sensitive data.

We have created a set of modules on Trailhead on “Building Ethical and Inclusive Products”, aiming to help our employees and the wider public learn how to create products with ethics, accessibility, and inclusion at their core.

**c. Does your organization communicate its risk management policies and practices with users and/or the public? If so, how?**

We make available detailed documentation on all our products.

We also recently published our first Trusted AI and Agents Impact report, in which we provide a comprehensive overview of our efforts and learnings in this area.

**d. Are steps taken to address reported incidents documented and maintained internally? If so, how?**

When we receive a report, our relevant teams track the reported incident, review it against our Acceptable Use Policy, and maintain review logs.

**e. How does your organization share relevant information about vulnerabilities, incidents, emerging risks, and misuse with others?**

We publish a range of resources on these topics that we have outlined in various sections of the reporting framework, including the publication of our security team on best practices for how organizations can protect themselves from AI security risks and vulnerabilities, such as prompt injections, data poisoning, and supply chain vulnerabilities.

We also participate in a number of government- and industry-led partnerships aiming to share information on these topics, including the U.S. National Institute of Standards and Technology (NIST).

**f. Does your organization share information, as appropriate, with relevant other stakeholders regarding advanced AI system incidents? If so, how? Does your organization share and report incident-related information publicly?**

On our trust.salesforce.com and salesforce.status.com websites, we provide transparency around service availability and performance for Salesforce products, including our AI products.

We also share resources on our online learning platform Trailhead, including on Artificial Intelligence and Risk Management; and Critical Incident Management at Salesforce.

We provide Security Advisories on our Security website to provide actionable information for customers.

**g. How does your organization share research and best practices on addressing or managing risk?**

Salesforce is a member of AISIC's Working Group #5: Safety & Security, which aims to coordinate and develop guidelines related to managing the safety and security of dual-use foundation models.

Salesforce is also participating in the OWASP Top 10 for LLM Core Team.

Salesforce is an active member in several Information Sharing and Analysis Centers (ISACs), dedicated to collecting, analyzing and disseminating actionable threat information between members.

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

We follow international best practices, for example by adhering to the Canada Voluntary Commitments on AI.

We have signed onto the EU AI Pact, agreeing to three key pledges:

- AI governance strategy to foster the uptake of AI and commitments to future compliance with the AI Act.
- Identifying AI systems likely to be categorized as high-risk under the AI Act.
- Promoting AI literacy and awareness among staff, ensuring ethical and responsible AI development.

Salesforce developers comply with industry standards including but not limited to OWASP 10, CWE25, NIST AI RMF.

**Any further comments and for implementation documentation**

*No answer provided*

## Section 5 – Content authentication & provenance mechanisms

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

Honesty is one of our trusted AI principles. This means respecting data provenance and ensuring consent to use data (e.g., open-source, user-provided) when collecting data to train and evaluate models. Additionally, transparency by design is key — when content is autonomously delivered (e.g., chatbot response to a consumer, use of watermarks), it is important to be transparent that an AI has created content.

We also view transparency and notification as critical design patterns for fostering trust in AI systems. This approach facilitates users' understanding: that they are informed whenever they interact with an AI system, comprehend the role of AI in generating outputs, and are aware of its capabilities and limitations. Notifications and disclosures, such as those embedded in Salesforce's AI agents, clarify whether communications or actions are AI-generated. This transparency builds user confidence by eliminating ambiguity, allowing people to make informed decisions about how they engage with the technology.

Notifications also play a key role in educating users about AI's impact and boundaries. For example, a disclaimer in an email generated by an AI system might explain that the content was auto-generated while highlighting that it adheres to ethical guidelines. Transparent design not only fosters accountability but also protects organizations from potential misuse or misunderstandings of AI applications. By making AI systems more comprehensible, this pattern bridges the gap between sophisticated AI capabilities and everyday user interactions.

Additionally, our AI Acceptable Use Policy mandates that customers disclose to their users when they are interacting with an AI feature, including Agentforce Agents.

**b. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems? If yes, how? Does your organization use international technical standards or best practices when developing or implementing content provenance?**

At the user interface level, we have implemented disclaimers for all of our AI experiences. It is required that all AI features or services include a disclaimer or user notification about the nature of generative AI content and its inherent risks, especially as we enter the era of agentic AI. And, our Audit Trail feature can be used by Salesforce administrators to see how the Einstein Trust Layer protects sensitive data from exposure to an external LLM and identify what content has been AI-generated.

At the product use policy level, as mentioned above, Salesforce's AI Acceptable Use Policy includes a requirement that customers disclose when an end user is interacting with AI and not mislead people into thinking AI-generated content was created by a human.

**Any further comments and for implementation documentation**

*No answer provided*

## Section 6 – Research & investment to advance AI safety & mitigate societal risks

**a. How does your organization advance research and investment related to the following: security, safety, bias and disinformation, fairness, explainability and interpretability, transparency, robustness, and/or trustworthiness of advanced AI systems?**

Salesforce's trusted AI principles and guidelines for responsible agentic AI position our technology as not only a powerful tool for business innovation but also a responsible, ethical partner in advancing societal goals and maintaining user trust.

Salesforce's AI Research team has published research on trust and safety metrics in LLMs and open-sourced a safety library, AuditNLG. We also have an internal Trust & Safety team responsible for proactive processes for reviewing, identifying, and mitigating product safety risks.

Our Research & Insights team within Product has released research about the importance of having the human touch in AI, particularly in an enterprise AI context.

Our Office of Ethical and Humane Use has funded ongoing user research to focus on trusted and responsible AI. So far, research topics have included: customer needs for watermarking and content provenance; customer needs for bias detection; end user needs for and concerns about autonomous agents; and Gen Z perceptions of AI — to name a few.

In Slack's Workforce Lab, studies have been conducted on the use of AI by studying the emotions affecting AI usage, how to promote prosocial AI, and the conditions that guide healthy AI usage habits.

**b. How does your organization collaborate on and invest in research to advance the state of content authentication and provenance?**

Our approach includes, among other things, continuous customer research. Our Research & Insights team has conducted research on customer expectations for watermarking and attribution.

Salesforce is an active member of the Data Provenance Initiative.

**c. Does your organization participate in projects, collaborations, and investments in research that support the advancement of AI safety, security, and trustworthiness, as well as risk evaluation and mitigation tools?**

Salesforce is a founding member of the WEF Center for Cybersecurity. Our Security team also participates in various ISACs (primarily FS-ISAC and RH-ISAC), and several other industry associations (e.g. CSA, OWASP). Each of these organizations is focusing on the implications of AI on Security as well as opportunities to improve capabilities using AI.

Our Office of Ethical and Humane Use team participates in the NIST Artificial Intelligence Safety Institute Consortium (AISIC); Singapore's AI Verify Foundation; and the WEF's Global Futures Council on Data Frontiers.

**d. What research or investment is your organization pursuing to minimize socio-economic and/or environmental risks from AI?**

To combat societal risks of bias and toxicity, we have implemented trust patterns in our AI aimed at addressing these issues. Examples include:

- Unchecked Demographics: By default, demographic attributes are unchecked when generated in marketing segments to help mitigate unintended bias.
- Model Containment: Prompt instructions are set to reduce the potential for toxic and biased outputs, including toxic mirroring. For example, a rule is built into the LLM so it won't use any words or phrases that are toxic, hateful, biased, inflammatory or offensive.

For our work on the environmental risks of AI, please see our responses under section 7 below.

**Any further comments and for implementation documentation**

*No answer provided*

## Section 7 – Advancing human and global interests

**a. What research or investment is your organization pursuing to maximize socio-economic and environmental benefits from AI? Please provide examples.**

Salesforce is committed to driving positive change for our employees, the environment, and society. Our focus on Employee Success, Sustainability, and Equality reflects our core values and our dedication to responsibly innovating with AI.

Salesforce's sustainable AI strategy, developed in collaboration between our AI Research, Sustainability, and Office of Ethical and Humane Use teams, focuses on optimizing models, utilizing energy-efficient hardware, and prioritizing low-carbon data centers to make our AI solutions as efficient and sustainable as possible. This strategy integrates sustainable practices at every stage, setting a high standard for responsible AI development.

Our Sustainable AI Policy Principles build on Salesforce's commitment to advocate for clear and consistent science-based policies for a just and equitable global transition to a 1.5˚C future. The principles offer clear best practices for lawmakers and regulators adopting sustainable AI regulations, including how to manage and mitigate the environmental impact of AI models and ideas to spur climate innovation with policies that can incentivize and enable the environmental application of AI.

Net Zero Cloud with Agentforce enables companies to streamline ESG management and reporting to comply with current disclosure frameworks. By integrating AI solutions with Net Zero Cloud, customers can manage and track ESG metrics, gain accurate and timely insights, streamline reporting, create compliant reports to achieve their ESG goals.

The Salesforce Accelerator – AI for Impact is a philanthropic initiative to help purpose-driven organizations gain equitable access to trusted generative AI technologies. The accelerator provides flexible funding, pro-bono expertise, and technology to nonprofits, empowering them to accelerate generative AI-based solutions to the world's most pressing challenges. Since 2023, the AI for Impact accelerator has helped 17 nonprofits build innovative AI solutions to advance equity in education and climate mitigation, adaptation, resilience, and finance.

Building on the success of the AI for Impact accelerator and the company's focus on agents, we launched the Salesforce Accelerator – Agents for Impact, an initiative designed to help nonprofits harness agentic AI. This accelerator will provide technology, funding, and expertise to help nonprofits build and customize AI agents, enabling them to improve operational efficiency and scale community impact in the AI-driven future.

We are enhancing accessibility by using AI to address specific needs across our products. For example, our Resize/Reflow initiative leverages AI to identify features that need to adapt to different screen sizes without losing context or function, and will eventually make interfaces more user-friendly, particularly for individuals who are blind or low vision.

In addition, our Self-Service for Engineering initiative supports product teams in addressing accessibility requirements proactively. Engineers and developers can access a repository of accessibility guidelines and answers to common questions proactively without waiting for a response from an accessibility engineer, ensuring accessibility considerations are integrated into product design and development from the start. This resource provides both foundational guidance and real-time support, making inclusive design more efficient and comprehensive across teams.

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

Trailhead, Salesforce's free online learning platform, has expanded its courses to offer AI-specific skills training, including AI fundamentals, ethical AI use and prompting.

Salesforce also offers its existing premium AI courses and AI certifications free of charge and available to anyone on Trailhead through the end of 2025.

Additionally, we make our spaces around the world available for on-site training sessions. In 2024, Salesforce opened its first AI Center in London, and unveiled Agentblazer Ranch at its headquarters in San Francisco in March 2025, with plans to roll out additional training centers in key hubs around the world like Chicago, Tokyo, and Sydney. These centers will bring together industry experts, partners, and customers to advance AI innovation alongside providing critical upskilling opportunities.

We also organize webinars for our customers on the risks and opportunities of AI, and publish relevant content regularly on our website.

To expand access to AI skills and careers for our employees, we created Career Connect, an internal AI-powered internal talent marketplace.

Salesforce philanthropic investments related to AI focus on championing a future where everyone benefits from AI equally. Our grantmaking in AI focuses on literacy and training, as well as tools and applications. In 2023, we gave $23 million to education to help the AI generation unlock critical skills. This funding included grants to U.S. school districts and global education nonprofits, including $6 million allocated to nonprofits focused specifically on AI skilling and literacy.

**c. Does your organization prioritize AI projects for responsible stewardship of trustworthy and human-centric AI in support of the UN Sustainable Development Goals? Please provide examples.**

Our Annual Stakeholder Impact Report includes a detailed overview of how we contribute to SGDs with the whole of our technology offering.

**d. Does your organization collaborate with civil society and community groups to identify and develop AI solutions in support of the UN Sustainable Development Goals and to address the world's greatest challenges? Please provide examples.**

We are an early member of the Coalition for Sustainable AI, through which we affirm our commitment to leverage the development and use AI to progress towards the UN Agenda 2030 and UN SDGs, focusing on those related to climate action and protection of the environment. The Coalition was initiated at the AI Action Summit in Paris in February 2025 by France, in collaboration with the UN Environment Program (UNEP) and the International Telecommunications Union (ITU).

In collaboration with Hugging Face, Cohere, and Carnegie Mellon University, Salesforce has released the AI Energy Score, a first-of-its-kind benchmarking tool that enables AI developers and users to evaluate, identify, and compare the energy consumption of AI models. Salesforce also announced it will be the first AI model developer to disclose the energy efficiency data of its proprietary models under the new framework.

We have also joined Current AI, a new global public interest AI partnership launched during the French AI Action Summit. Current AI will develop and support large-scale initiatives that serve the public interest. Key focus areas include healthcare, linguistic diversity, science, and issues such as trust and safety, and AI auditing.

We also address societal impacts through our investments and community engagements. Examples include our investments in initiatives like Salesforce's climate accelerator, education accelerator, and AI for impact accelerator.

**Any further comments and for implementation documentation**

*No answer provided*