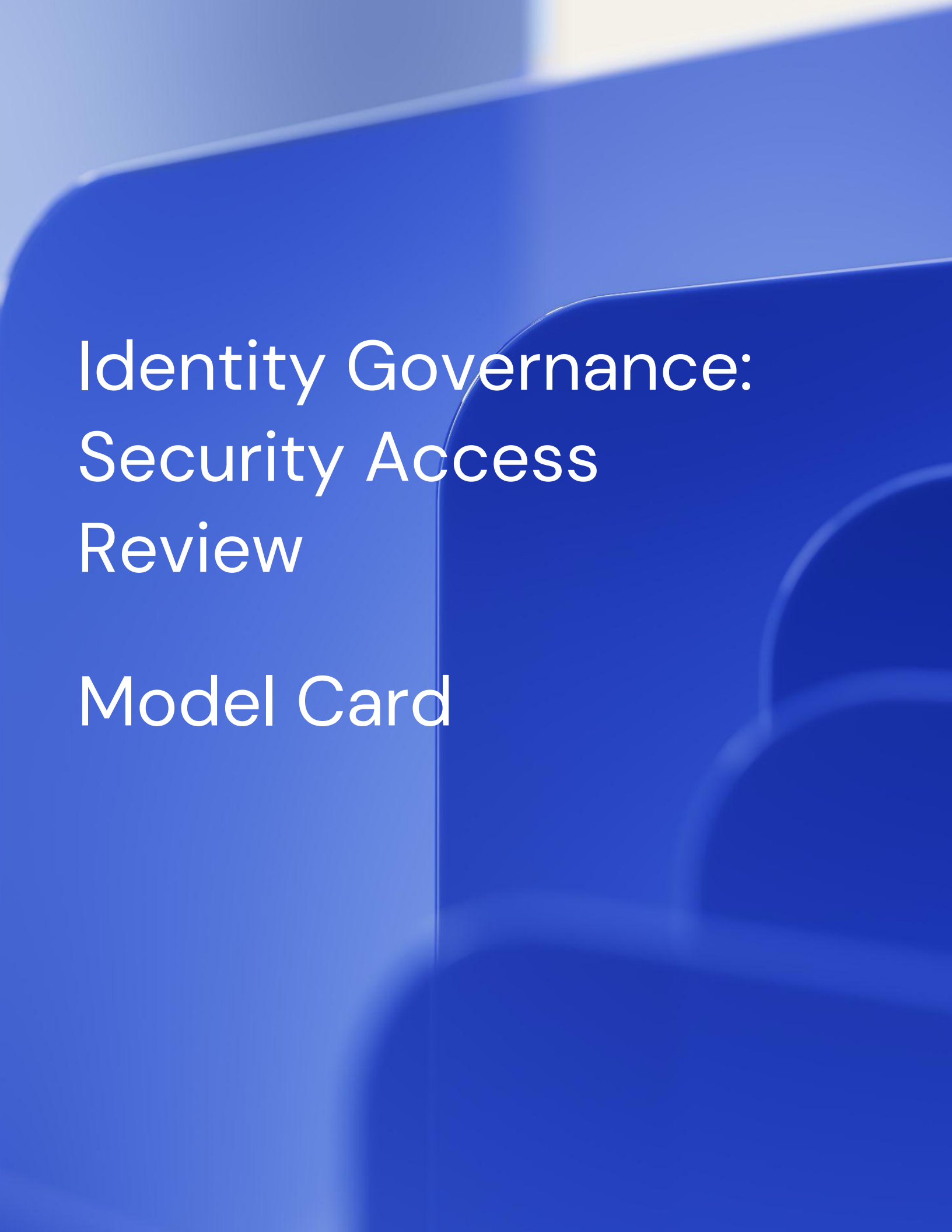

Identity Governance: Security Access Review

Model Card

Okta Model Cards are intended to provide information about models leveraged by Okta in Okta's products and services and include information on the intended use cases, limitations, training, and evaluation of models. Model cards are not intended to be technical reports and are provided for informational purposes only. Model cards may be updated from time-to-time.

Model Card: Okta Identity Governance: Security Access Review

Overview

- **Product/Feature Name:** Okta Identity Governance: Security Access Review
- **Availability:** Security Access Review is currently available in GA.
- **Description:** The Security Access Review feature is aimed at providing customers with a unified security platform which empowers administrative users to quickly understand, respond to and remediate potential user risks related to end user application/resource access. The primary output will be risk-related summaries of an application or user's access to applications to provide administrative reviewers with information needed to determine whether certain users should have continued access.
- **Primary Function:** Analysis, Insights, and Content Summarization/Description

---

Model Details

- **Model Type:** Large language model (LLM)
- **Model Origin:** Third Party
- **Model Provider:** Anthropic, hosted on AWS Bedrock. The model is hosted by AWS through Amazon Bedrock under an escrow arrangement, meaning the models do not run on Anthropic's infrastructure.
- **Specific Model Name/Version:** Claude Sonnet 3.7
- **How is the model accessed?** Managed Service, specifically Amazon Bedrock.

Intended Use & Limitations

- **Intended Use Cases:** Summarize potential risk-related information regarding a user's access to applications in order to assist human administrators in their review process.
- **Out-of-Scope Use Cases:** Any use other than the intended use case is out of scope and not recommended.
- **Known Limitations:**
    - The models can sometimes provide responses that sound plausible but are inaccurate or incomplete. Always verify the accuracy of the responses.
    - The quality of the responses may depend on various factors, including but not limited to, the complexity and scale of the user's access. Okta recommends that customers review and assess the output for the customer's specific use case. Any information provided by this feature is for informational purposes only and does not constitute legal, privacy, security, compliance or business advice.

- **Potential Risks:** What are the potential risks or ways the model could fail or produce problematic outputs? Check all that apply and briefly explain:

  Factual Incorrectness (Hallucinations): There is risk that the summaries could be inaccurate, e.g., that certain system access is identified incorrectly as riskier or not as risky. This risk increases with a larger amount of data (e.g., if a user has complicated entitlements, or access to many apps).
  Bias: The model may produce outputs that are biased against certain demographic groups or reflect societal stereotypes.
  Harmful or Inappropriate Content: The model could generate offensive, unsafe, or otherwise inappropriate content.
  Other: [[e.g., susceptibility to prompt injection, generating insecure code]

Data

- **Model Inputs:** Access Governance data (e.g. which users have access to applications and resources, how such access was granted, etc.) in JSON format.
- **Model Outputs:** Summaries of a user's access to applications/resources and whether that access should be continued.
- **Data Minimization:** Guardrails are in place to filter the input and output of the LLM for personal information.
- **Training Data:** The model is not trained by Okta, it is trained by the model provider.
- **Is the model trained on Customer Data** (as defined in Okta's Master Subscription Agreement at https://www.okta.com/legal)**?** The model does not train on Customer Data.

Evaluation

- **Methodology:** The performance is evaluated using both an LLM as a judge and human evaluation as needed. In addition, guardrails are applied at the model serving layer to help reduce risks such as harmful or irrelevant outputs, including measures for prompt filtering, output validation, and content moderation.
- **Performance Metrics:** Metrics used include correctness, completeness, faithfulness, helpfulness, logical coherence, relevance, following instructions, and professional style and tone. Okta does not publish specific performance metrics, as this information could be exploited by attackers to understand the model's strengths and weaknesses.

Artificial Intelligence (AI) Principles

Okta strives to safely use and develop AI to strengthen the connections between people, technology, and our community. When it comes to AI innovation, we aim to live our core values and harness the power of AI in a way that reflects said values. This kind of thinking is part of our DNA. That's why we take a values-based approach to AI. Okta's Responsible AI Principles underscore (i) transparency; (ii) building customer trust through security, privacy, and safety; (iii) accountability; and (iv) innovating responsibly regarding inclusivity, fairness, and ethics.

These principles are aligned with Okta's values: "Love our customers." "Always secure. Always on." "Build and own it." "Drive what's next."

Our developers adhere to responsible AI principles regarding privacy, security, responsible innovation, and more general principles and obligations regarding Customer Data. For more information, please see the published full version of Okta's Responsible AI Principles on Okta.com.

Security and Privacy

- Okta adheres to its existing commitments regarding security, privacy, and confidentiality in connection with Okta products and features that leverage AI that are offered as part of the Okta services.
- Okta follows industry standard processes for testing, developing, and making available products and features that leverage AI for customers.
- Okta has policies and programs in place regarding the use of and governance over AI.
- The data validation measures Okta takes for products and features that leverage AI may vary by product and feature and may include measures like input sanitization, having an allow list of characters that can be passed in the input, having a block list of terms that will be rejected, and having a custom post processing step that validates the output depending on the use case.
- The measures Okta has in place to help ensure that the models leveraged by Okta in Okta's product offerings are accurate and unbiased may vary by product and feature and may include monitoring the performance of models, auditing data to identify inaccuracies or missing information, having a diverse team of developers and data scientists that develop, maintain and improve Okta's products that leverage AI, and having a human in the loop when necessary.

Last Updated: December 19, 2025