Auth0 Security AI Agent

Model Card

Okta Model Cards are intended to provide information about models leveraged by Okta in Okta's product offerings 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: Auth0 Security AI Agent

Overview

- **Product/Feature Name:** Auth0 Security AI Agent (Beta)
- **Availability:** Auth0 Security AI Agent is currently only available in beta.
- **Description:** The Auth0 Security AI Agent is a collection of AI agents that uses a Large Language Model (LLM) to assist Auth0 tenant administrators in responding to security anomalies by conducting triage investigations, and recommending actions for tenant anomalies. All recommended actions require explicit user confirmation, providing an ongoing human validation layer for model output quality.
- **Primary Function:** Content Generation, Analysis & Insights, and Process Automation.

Model Details

- **Model Type:** Large Language Model (LLM)
- **Model Origin:** Third-Party Model
- **Model Provider:** Anthropic
- **Model Version:** Claude Sonnet 4.5/4.6 or Claude Opus 4.5/4.6/4.7
- **How is the model accessed?** Managed Service, specifically Amazon Bedrock.

Intended Use & Limitations

- **Intended Use Cases:** The Security AI Agent is designed to assist Auth0 tenant administrators by performing triage investigations, and recommending actions for detected tenant anomalies. This feature is only available to tenant admins.
- **Out-of-Scope Use Cases:** The following are excluded from the beta and are not supported at this time:
    - Slow burn attacks do not trigger rate limits, which is our only detection method in Beta.
    - Attack types beyond credential stuffing and brute force.
    - Free-form natural language queries (Beta is button-driven only)
    - **Important: All recommended actions require proper review and approval by the user before application. No recommended actions are automatically executed by the Security AI Agent.**

- **Known Limitations:**
    - The model requires at least 28 days of daily data and 72 hours of hourly data to establish reliable baselines.
    - Tenant Data inputs are limited to a specific set of predefined fields, including Tenant ID, JA4, IP, User-Agent (UA), GeoLocation, Time anomaly was first detected, Rate Limit (per second and count), Tenant Logs counts, and this data is shown to the model in a single pre-filtered format.
    - Novel attacks requiring actions not on the predefined list receive suboptimal recommendations.
- **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):** The model may generate information that is not factually correct.
    - ☑ **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:** Prompts for various agents; Auth0 Playbooks (scraped text); and Tenant Data, including Tenant ID, Rate Limit, Time anomaly was first detected, Rate Limit (per second and count), Tenant Logs counts, Top IP, JA4, UA, and GeoLocation data.
- **Model Outputs:** The model produces the following output types:
    - Triage Investigation Reports
    - Security Recommendations
- **Data Minimization:** The model uses the tenant data necessary to perform triage investigation and generate recommendations.
- **Training Data:** The model was not trained by Okta. It was 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)**?** No.

Evaluation

- **Methodology:** The model was evaluated using the "LLM-as-a-judge" RAGAS framework. Model performance and operational health are continuously monitored.
- **Performance Metrics:** Performance monitoring metrics include token usage, time to response, and user sentiment feedback.

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 June 3, 2026