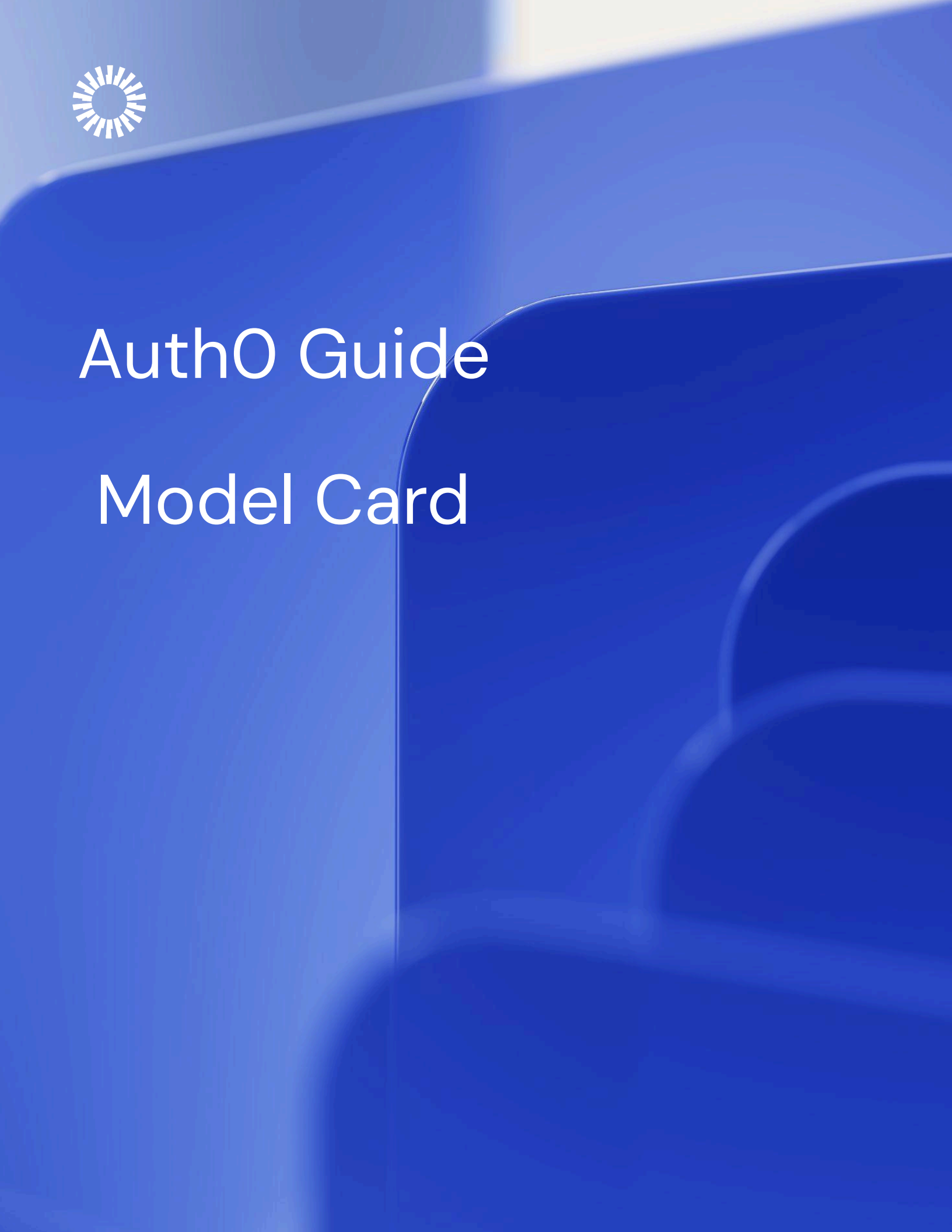

Auth0 Guide

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: Auth0 Guide

Overview

- **Product/Feature Name:** Auth0 Guide
- **Availability:** Auth0 Guide is currently only available in early access[1].
- **Description:** The Auth0 Guide is a chatbot located in the Auth0 Management Dashboard. It helps users by answering questions about Auth0 using information from Auth0 documentation, blog posts, and community forums that have verified answers. Two artificial intelligence (AI) models are used: one vector embedding model that finds the relevant information for your query and two generative AI (GenAI) models, specifically large language models (LLM), that analyze and summarize the information into a clear and concise answer.
    - Auth0 Guide provides Enterprise customers with the additional capability to query the customer's aggregated Security Center data to obtain information, such as the number of logins or threats detected within the time period requested by the customer, which cannot exceed the last 30 days.
        - Enterprise customers who have purchased the Attack Protection add on will have the additional capability to obtain certain metrics.
- **Primary Function(s):**
    - For the Vector Embedding Model: Search and Retrieval
    - For the LLM: Analysis and Insights and Content Summarization

---

Model Details

- **Model Types:** Three AI models are used: one vector embedding model and two LLM.
- **Model Origin:** Third-party models.
- **Model Provider:** The LLMs are provided by Anthropic and the Vector Embedding Model is provided by Cohere. The models are hosted by AWS through Amazon Bedrock under an escrow arrangement, meaning the models do not run on Anthropic's or Cohere's infrastructure.
- **Specific Model Name/Version:** Anthropic Claude Sonnet 4.0 and Cohere Embed V3 English.
- **How is the model accessed?** Managed Service, specifically Amazon Bedrock.

---

[1] Any products, features, functionalities, certifications, authorizations, or attestations referenced in this material that are not currently generally available or have not yet been obtained or are not currently maintained may not be delivered or obtained on time or at all. Product roadmaps do not represent a commitment, obligation or promise to deliver any product, feature, functionality, certification or attestation. Do not rely on them to make purchasing decisions.

Intended Use & Limitations

- **Intended Use Cases:**
    - Auth0 Guide is designed to answer questions about Auth0.
    - Auth0 Guide's capability to access and query the Customer's aggregated Security Center data is intended to provide Enterprise customers with additional information, such as data analysis, recommendations for protecting Auth0 tenants, support for reviewing Security Center alert notifications, and metrics, if applicable.
- **Out-of-Scope Use Cases:**
    - The models are not designed to answer questions about pricing;
    - The models are not intended for use in high-risk or safety-critical systems; and
    - 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 clarity and complexity of questions. For best results, ask specific questions about a single topic. Guide may not perform as expected when asked about complex, multi-part, or highly nuanced topics.
    - For Auth0 Guide's capability to query the customer's Security Center data, Guide's knowledge of Security Center data is currently limited for complete incident management, and any recommendations for protecting Auth0 tenants are based on generally accepted industry best practices, Okta's own standards, and the models' underlying training and knowledge, but may not take into consideration all of a customer's specific circumstances. Okta recommends that customers review and assess the recommendations for the customer's specific use case. Any recommendations provided by this feature are for informational purposes only and do not constitute legal, privacy, security, compliance or business advice.
- **Potential Risks:** What are the potential risks or ways the models could fail or produce problematic outputs? Check all that apply and briefly explain:
    - ☑ **Factual Incorrectness (Hallucinations):** The models 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:** The models receive text strings from the user (questions or statements), internal system prompt(s), documentation, Security Center data, and the user's timezone.
- **Model Outputs:** Output generated text as an answer to the user's question.

- **Data Minimization:** Auth0 Guide is designed for general use and should not be used to process personal data or sensitive data. The applicable documentation for Auth0 Guide advises against the submission of any personal or sensitive information to the model.
- **Training Data:** The models are not trained by Okta, they are trained by the model providers.
- **Is the model trained on Customer Data** (as defined in Okta's Master Subscription Agreement at https://www.okta.com/legal)**?** The models do not train on Customer Data.

Evaluation and Security

- **Methodology:** The performance evaluation includes 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:** 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 October 23, 2025