



Elastic Trust Center — Product AI — Frequently Asked Questions



About this document

This FAQ addresses the most common security, privacy, and governance questions received from customers and prospects regarding Elastic's AI capabilities. It is organized by AI capability type rather than deployment type, reflecting the way enterprise customers and regulators evaluate AI risk. For deployment-specific considerations (Elastic Cloud Hosted, Elastic Cloud Serverless, Elastic Self-Managed, and standalone air-gapped environments), please refer to the relevant notes within each section.

Elastic AI capabilities covered: Elastic AI Assistant (Security and Observability), Elastic Inference Service (EIS), Elastic Agent Builder, built-in embedding and reranking models, anomaly detection (ML), Jina On-Prem (commercially licensed Jina AI models for self-managed and air-gapped deployment), and customer-configured third-party large language model (LLM) connectors.

Section 1: AI governance and program oversight

Applicable to all Elastic AI capabilities and deployment types

Q1: Does Elastic have a formal AI governance program?

Yes. Elastic operates a formal AI governance program aligned to the ISO/IEC 42001:2023 AI Management Systems framework, which is currently in progress. All AI tools and use cases are subject to review and approval by a multi-disciplinary review team comprising InfoSec, Legal, IT, and Engineering. This team evaluates performance quality, training data provenance, legal and regulatory risk, and safety features before any AI capability is approved for use. A registry of AI systems in use is maintained, and AI roles and responsibilities are formally defined.

Q2: What is Elastic's certification and compliance posture for AI?

Elastic's AI features are delivered within an infrastructure certified to ISO/IEC 27001:2022, ISO/IEC 27017, ISO/IEC 27018, ISO/IEC 27701:2019, SOC 2 Type II, PCI DSS v4.0, and HIPAA-eligible configurations (for Elastic Cloud Hosted and Serverless). Elastic is actively pursuing ISO/IEC 42001:2023 certification. Elastic's AI use cases have been assessed as not "High Risk" under the EU AI Act. However, Customers remain responsible for



evaluating whether their custom workflows or specific operational uses trigger additional High-Risk requirements. The AI features delivered via Elastic Inference Service and AI Assistants fall within the security control scope of these certifications.

Q3: Does Elastic use customer data to train its AI models?

No. Elastic does not use customer data — including prompts, queries, responses, or ingested content — to train, fine-tune, or improve its AI models. Elastic's built-in embedding and reranking models were trained exclusively on public and open source data. When customers use Elastic's managed inference service (EIS), Elastic selects providers operating under zero-data-retention agreements, meaning prompts and responses are not used for training by those providers either.

Q4: How does Elastic assess and manage risk for new AI capabilities?

All new AI tools and use cases undergo a formal review before approval, considering data sourcing, known vulnerabilities, bias risk, and applicable safety features. AI system impact assessments are conducted for Elastic-developed models. Elastic's AI governance program includes a formal Statement of Applicability covering all Annex A controls under ISO/IEC 42001, documented AI policies, risk assessment processes, and defined roles and responsibilities. Third-party AI providers are additionally evaluated for security, compliance, and safety before being onboarded to Elastic's managed inference service.

Q5: What is Elastic's EU AI Act risk classification?

Elastic has assessed its AI features as not "High Risk" under the EU AI Act. Elastic AI Assistant and Elastic Inference Service are general-purpose tools for search, security, and observability use cases — they do not fall into the EU AI Act's enumerated high-risk categories, such as employment decision-making, biometric identification, access to essential services, or critical infrastructure control. All AI outputs are advisory, with human oversight required before any consequential action is taken. Because legal classification depends on the final operational context, customers remain responsible for their specific implementations. If a customer integrates Elastic AI into custom workflows that fall within high risk categories under Annex III of the EU AI Act, the customer acts as the AI deployer under the Act and is responsible for fulfilling deployer compliance duties.



Q6: What AI models does Elastic offer, and can customers choose their own?

Customers have three options:

- Use Elastic-managed LLMs via EIS: Elastic's Inference Service provides access to models from providers including Anthropic, OpenAI (GPT), Google (Gemini), Meta, and Amazon. All supported models are listed at elastic.co/docs/explore-analyze/elastic-inference/eis-supported-models.
- Bring your own LLM: Customers may configure their own third-party LLM connectors (e.g., Azure OpenAI Service, Amazon Bedrock, OpenAI, Google Vertex). In this case, the model and all associated data practices are governed by the customer's agreement with that provider.
- Opt out: Customers may choose not to use generative AI features. Elastic also offers non-generative AI models—ELSER, E5, and Elastic Rerank—that run entirely within the customer's cluster.

Section 2: Elastic's built-in ML models

Embedding, reranking, and anomaly detection models running within the customer's cluster

Q7: What are Elastic's built-in ML models and where do they run?

Elastic provides proprietary, non-generative machine learning models for tasks including semantic search (embedding), relevance improvement (reranking), and anomaly detection. These models run entirely within the customer's Elasticsearch cluster and do not make external network calls. No customer data leaves the Elastic environment as a result of using these models. They are available across Elastic Cloud Hosted, Elastic Cloud Serverless, and self-managed deployments.

Q8: How are Elastic's built-in models trained, and is customer data used?

Elastic's built-in models are trained on public and open source data only. Customer data is never used to train, validate, or fine-tune these models. Model performance is benchmarked publicly against standard retrieval-accuracy benchmarks, and bias and fairness are assessed as part of the model development process under Elastic's AI governance framework.

**Q9: How are Elastic's built-in ML models secured and updated?**

Built-in ML model artifacts are distributed as part of Elastic's standard software release process and are subject to the same vulnerability management and patch management procedures that apply across Elastic's products. Model updates are delivered through standard product upgrades. For Elastic Cloud Hosted and Serverless deployments, Elastic manages and applies model updates. For self-managed deployments, the customer is responsible for applying product updates.

Section 3: Elastic Inference Service (EIS) — Elastic-managed LLM gateway

Elastic's managed gateway providing access to vetted third-party large language models

Q10: What is Elastic Inference Service (EIS) and how does it work?

Elastic Inference Service (EIS) is Elastic's managed gateway that routes generative AI inference requests to a curated set of vetted third-party large language model (LLM) providers, including OpenAI, Anthropic, Amazon, Google, and Meta. EIS is available as a fully integrated, zero-configuration service on Elastic Cloud Serverless, and as an opt-in service on Elastic Cloud Hosted and self-managed deployments (from Elastic Stack 9.0+ with an Enterprise license via Cloud Connect). All EIS providers operate under zero-data-retention agreements with Elastic, meaning inference payloads are not stored or used for model training by those providers.

Q11: What data is transmitted to EIS providers, and is it retained?

When a customer uses EIS, inference request payloads (prompts and context) are transmitted to the relevant LLM provider over mTLS/TLS 1.2+ for processing. These payloads are not retained at rest by Elastic or, per zero-data-retention agreements, by the LLM provider after the request is fulfilled. Only non-content usage metadata — model identifier, token counts, organization ID, and timestamps — is retained by Elastic for metering and billing purposes. Authorization headers and prompt content are explicitly excluded from EIS logs.



Q12: How does Elastic vet and manage its third-party LLM providers?

Elastic evaluates candidate LLM providers for performance, accuracy, security, safety, and legal and regulatory compliance before onboarding them to EIS. Onboarded providers are required to operate under zero-data-retention agreements with Elastic. Elastic's vendor management program includes risk classification, security requirements, and at least annual review of critical vendor attestation reports. The current list of supported EIS providers and models is published at elastic.co/docs/explore-analyze/elastic-inference/eis-supported-models.

Q13: How is EIS traffic secured in transit?

All inference traffic through EIS is encrypted using TLS 1.2 or higher. Client connections to the EIS Gateway use mutual TLS (mTLS). Connections from the EIS Gateway to upstream LLM providers are protected by TLS 1.2+. Secrets and credentials for EIS infrastructure are managed via a centralized secrets vault with pre-push secret scanning. Inference traffic is also scoped per tenant using organization- and account-level access controls.

Q14: Has penetration testing been performed on the EIS and AI features?

Yes. Elastic has conducted penetration testing covering EIS and related AI features, including automated and manual prompt injection and jailbreaking attempts. Findings are tracked through Elastic's standard vulnerability management process. A penetration test report and addendum are available for qualified customers via Elastic's assurance profile at assurance.elastic.co.

Section 4: Customer-configured third-party LLM connectors

Customer-managed connections to external LLMs (e.g., Azure OpenAI, Amazon Bedrock)

Q15: Can customers connect their own LLM to Elastic's AI features?

Yes. Customers may configure their own third-party LLM connectors — for example, Azure OpenAI Service, Amazon Bedrock, or any compatible endpoint — as an alternative



to Elastic's managed EIS. In this configuration, inference traffic flows directly from the customer's Elastic deployment to their chosen LLM provider under the customer's own agreement. Elastic has no visibility into this traffic, and the data practices and security controls for that path are entirely governed by the customer's contract with their LLM provider.

Q16: Can customers restrict which LLMs are accessible within their Elastic environment?

Yes. Customer administrators control which LLM connectors, if any, are configured within their Elastic environment. Customers may restrict AI features to a specific approved connector, limit which users or roles can invoke AI capabilities via RBAC, or opt out of generative AI entirely and use only Elastic's non-generative ML models. Network egress controls can additionally be configured to restrict AI traffic to approved endpoints.

Section 5: Elastic AI Assistant and agentic capabilities

Elastic AI Assistant (Security and Observability) and Elastic Agent Builder

Q17: Does Elastic's AI Assistant take autonomous actions on production systems?

No. The Elastic AI Assistant (available in Security and Observability) is designed as an advisory tool. It presents outputs — summaries, recommended queries, suggested actions, and enriched context — to human operators for review. User confirmation is required before any suggestion is acted upon, and the Assistant does not autonomously execute actions on production systems. Elastic's Agent Builder capability can be configured by customers to enable agentic workflows, but the scope and permissions of any automated actions are defined and controlled by the customer, and agent access is limited to data the requesting user is authorized to access.

Q18: How does Elastic protect against prompt injection in the AI Assistant?

Elastic tests for prompt injection as part of third-party adversarial testing engagements, with findings tracked through standard vulnerability management. The AI Assistant



supports configurable data anonymization prior to sending context to the LLM — customers can define anonymization rules to redact sensitive fields (such as IP addresses or usernames) from the context window before it is transmitted. Any prompt injection weaknesses identified through testing are addressed under Elastic's Secure Software Development Framework (SSDF).

Q19: How is conversation history managed in the AI Assistant, and can it be deleted?

Conversation history persistence in the AI Assistant is disabled by default. If enabled by a customer administrator, conversation records are stored in a customer-controlled Elasticsearch index within the customer's own deployment — not in Elastic's infrastructure. Customer administrators can delete individual conversation records or purge all AI interaction history via Elasticsearch index management APIs or the Kibana data management interface. Elastic supports data deletion with written attestation upon contract termination.

Section 6: Jina On-Prem: Air-gapped and self-managed AI models

Elastic's Jina AI models, commercially licensed and packaged for on-premises and air-gapped deployment

Q20: What is Jina On-Prem?

Jina On-Prem is a commercially licensed packaging of Elastic's Jina AI models, including reader, embedding, and reranking models, that customers run entirely within their own infrastructure, including on-premises and air-gapped environments. The models are distributed as Docker containers that customers download and install into infrastructure they control. Jina On-Prem is designed for organizations that cannot, or choose not to, send data to cloud-based AI services, including public sector, defense, financial services, and healthcare customers.



Q21: Does Jina On-Prem send any data outside the customer's environment?

No. Once deployed, Jina On-Prem makes no outbound network calls: there is no license server, no telemetry or logging endpoint, and no connection to a model registry or to Hugging Face. All inference is performed within the customer's own environment, and customer data, including documents, queries, and generated embeddings, remains entirely within infrastructure the customer controls. Elastic has no access to, or visibility into, data processed by a customer's Jina On-Prem deployment.

Q22: What models are included in Jina On-Prem, and how were they trained?

Jina On-Prem includes the Jina AI model family covering reader (data extraction), embedding, and reranking models, spanning text, image, audio, and video modalities. As with all Jina models, these are trained on publicly available data; customer data is never used to train, fine-tune, or validate them. Model quality is evaluated against public industry benchmarks (such as MTEB and BEIR) as part of Elastic's model development and release process.

Q23: How does Jina On-Prem licensing differ from the Jina models on Hugging Face?

Jina models published on Hugging Face are licensed for noncommercial use (primarily CC BY-NC 4.0), which permits testing and evaluation but not commercial production deployment. Jina On-Prem provides a commercial license for running the models in production within self-managed, on-premises, and air-gapped infrastructure, backed by Elastic. For licensing details, contact your Elastic representative.

Q24: Does Jina On-Prem require Elasticsearch or Elastic Inference Service?

No. Jina On-Prem runs independently and does not require Elasticsearch or any connection to Elastic Cloud or Elastic Inference Service (EIS). It exposes standard API schemas compatible with common embedding APIs, and it can serve as a drop-in replacement for Jina models served through EIS — allowing air-gapped Elastic deployments to use the same models without changing how their applications call embedding and reranking endpoints.



Q25: Who is responsible for securing and operating a Jina On-Prem deployment?

Because Jina On-Prem runs entirely within the customer's environment, the customer is responsible for the security of the underlying infrastructure, network isolation, access controls, and monitoring, consistent with a self-managed deployment model. Elastic is responsible for the development, evaluation, and release of the model artifacts and container images.

Deployment note: Jina On-Prem is a separate offering from Elastic's built-in ML models (Section 2) and from Jina models served via EIS (Section 3). It applies to self-managed and air-gapped environments only.