Elastic Managed LLM

Intended Use and Functionality

Purpose of the model
The Elastic Managed LLM is intended to be used for generative AI features that are part of the Elastic solutions - Search, Security and Observability. The chosen model performs well with the Elastic AI assistants (Search, Security and Observability), and other AI features such as Attack Discovery, Automatic Import, Automatic Migration and others.

User Benefits
The Elastic Managed LLM is intended to be used within the Elastic platform for AI features in our three solutions - Search, Security and Observability. Elastic is building AI-enabled features across its platform and solutions and each feature that requires the use of LLMs or other non-generative models will provide documentation and specific instructions where applicable.

Model Architecture
We are currently using Anthropic Claude 3.5 Sonnet as the model for the Elastic Managed LLM hosted on Amazon Bedrock. For details on the model refer to the Claude 3.5 Sonnet model card.

Technical Architecture
A Customer project or deployment hosted in any CSP/Region will have access to the Elastic Managed LLM hosted in AWS US regions. All data is encrypted in transit. We configure a zero retention policy with the third-party LLM i.e. none of the AI inputs or outputs are retained by the LLM.

Optimization scope and limitations
We tune our solutions, optimizing prompt structure and context windows to work best with the AI Assistants (Search, Security and Observability), Security and Observability specific use cases such as Attack Discovery, Automatic Import and others.

As of now, Claude 3.5 Sonnet on Amazon Bedrock is one of the best performing models with the region availability and security characteristics we are comfortable providing customers with.

Risks
Given that Claude 3.5 Sonnet is the model for the Elastic Managed LLM, we carry the risks associated with the third party LLM. For more details read the Section 3.2 Safety Evaluations Overview in the Claude 3.5 Sonnet model.

Ethical considerations
Developers should not use the Elastic Managed LLM to generate malware or in any other way that is prohibited by its usage restrictions according to Claude's family of models on Amazon Bedrock. Elastic does not endorse or assume any responsibility for the accuracy, completeness, legality, or appropriateness of the results generated by such tools. Please report instances of

hallucinations, malicious code, or unwanted data in output so that we can evaluate for remediation to support@elastic.co.

Training or Fine Tuning
Elastic does not do any further pre-training or fine-tuning on this model and the model is used as-is, as provided by Claude 3.5 Sonnet on Amazon Bedrock.

Evaluation data
We evaluated the performance of Claude 3.5 Sonnet and several other models for Elastic use cases. We have published a model performance matrix based on our internal evaluation for all security use cases.

For the Observability AI assistant, we are currently evaluating several scenarios. The scenarios include tests related to alerts, APM, documentation, ES related functions, ES|QL queries and knowledge base related functions. We do not currently publish a model performance matrix for Observability and Search.

Technical Means for Integration
All data is encrypted in transit. We have a zero retention policy set for Amazon Bedrock, which means no AI input or output is stored in Amazon Bedrock. We expect customers to exercise the same caution with the Elastic Managed LLM as they do with any third-party LLM.