General Purpose LLM v2

Intended Use and Functionality

Purpose of the model

The General Purpose LLM v2 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.

Model Architecture
We are currently using Anthropic Claude 4.5 Sonnet as the model for the General Purpose LLM v2 hosted on Amazon Bedrock. Customers located outside the US are advised that enabling this feature requires transferring data to the US for processing during inference, even if no data is retained. For details on the model, refer to the Claude 4.5 Sonnet model card.

Technical Architecture
A Customer project or deployment hosted in any CSP/Region will have access to the General Purpose LLM v2 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.

Risks
The inherent risks of the underlying third-party model (Claude 4.5 Sonnet), such as potential hallucinations or bias, remain applicable.
For more details, read Section 2 Safeguard and harmlessness in the Claude 4.5 Sonnet model.

Ethical considerations
Developers should not use the General Purpose LLM v2 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 4.5 Sonnet on Amazon Bedrock.

Evaluation data
We evaluated the performance of Claude 4.5 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 General Purpose LLM v2 as they do with any third-party LLM.