

Webex AI Agent

AI Transparency Technical Note

At Cisco, we harness the power of artificial intelligence (“AI”) to drive technological innovation and deliver exceptional value to our customers. We also recognize that by applying this technology, we have a responsibility to mitigate potential harm. That is why Cisco adheres to our Responsible AI Framework ([“RAI Framework”](#)), which is based on six principles: Transparency, Fairness, Accountability, Privacy, Security and Reliability (the [“RAI Principles”](#)).

To that end, our teams build features that leverage AI with transparency, fairness, accountability, privacy, security and reliability at their core. To ensure that AI in Cisco Offers aligns with our RAI Framework and RAI Principles, Cisco Offers that include AI are scrutinized through our AI Governance process.

This AI Transparency Technical Note (ATTN) describes more information about Webex AI Agent and how we responsibly leverage AI technology to innovate and unlock new possibilities.

Contents

1. Feature Overview	3
2. Model Overview	4
Introduction	4
Usage Guidelines	6
Model Inputs and Outputs	7
3. Data Sources for Training & Education	9
Data Locality for AI Agents Components	9
4. Model Evaluation and Performance	11
5. Safety, Accuracy, and Fairness	12
Safety and Accuracy	12
Fairness	12
6. Privacy and Security	13
7. Monitoring and Maintenance	14
8. References	15

1. Feature Overview

Webex AI Agent is a platform that is designed to create, manage, and deploy automated AI agents to fulfil customer service and support needs. Using artificial intelligence, AI Agents provide automated assistance to customers before they interact with human agents over voice or digital channels.

These AI agents can function in either an **autonomous** mode or a **scripted** mode. While autonomous agents leverage Large Language Models (LLMs) for driving the dialog and state management, scripted agents use conventional machine learning algorithms for Natural Language Understanding (NLU) to capture user intents and respond as per configuration. Scripted agents with Webex AI Pro 2.0 incorporates a Large Language Model (LLM) to enhance intent classification accuracy.

Both autonomous and scripted AI agents support multilingual capabilities, enabling them to understand and communicate in multiple languages, enhancing accessibility and user experience across diverse linguistic backgrounds.

2. Model Overview

Introduction

Autonomous AI agents

Autonomous AI Agents are powered by sophisticated AI engines that combine speech technology (Speech-to-Text/Automatic Speech Recognition and Text-to-Speech),

Large Language Models (LLMs), intelligent guardrails, and expertly crafted system prompts.

Key Components of AI Engine:

- **Large Language Model:** Powers the AI Agent with its advanced intelligence, enabling it to understand complex queries, generate coherent responses, perform actions or answer from a knowledge base.
- **Guardrails:** Enable responsible AI interactions by setting clear boundaries, preventing inappropriate content, and maintaining brand safety.
- **Speech-to-Text (ASR/STT):** Converts spoken language into text, enabling AI agents to understand human speech.
- **Text-to-Speech (TTS):** Converts text into natural, human-like speech, so the AI agent can communicate clearly and engagingly.
- **Interim Response Model:** Provides real-time responses for user queries, creating a highly responsive and fluid user experience.
- **Standalone Query Generator Model:** transforms non-contextual, incomplete or colloquial inputs (e.g., “interest rate”) into complete, self-contained queries (e.g., “What is the interest rate for home loans?”), enabling more accurate retrieval and response generation.
- **Turn Detection Model:** Intelligently anticipates when a user has finished speaking, facilitating natural conversational flow.

The LLMs that power the Autonomous AI agents are from Microsoft's Azure OpenAI Service. These agents can utilize the following underlying AI engines: Webex AI Pro 1.0, powered by Azure for Text-to-Speech (TTS) and Speech-to-Text (STT), Webex AI Pro-US 1.0, and Webex AI Pro-EU 1.0, which uses third-party speech providers such as Eleven Labs for TTS and Deepgram for STT. For more details on Azure OpenAI Service and how it handles data, please review [Microsoft's transparency note available here](#).

1. The Webex AI Agent platform provides additional instructions, guardrails and context to the Azure OpenAI models to ensure a well-orchestrated experience for the users. The solution currently uses the GPT-4o, GPT-4.1-mini, GPT-4.1 and GPT-4.1-global ag models (for variant generation and other features) offered by Azure OpenAI Service. Based on our research and benchmarking, we found these models to provide the best experience for conversational AI use cases in terms of cost, latency, and quality. However, we are constantly evaluating available options and may choose to switch to or offer other models.

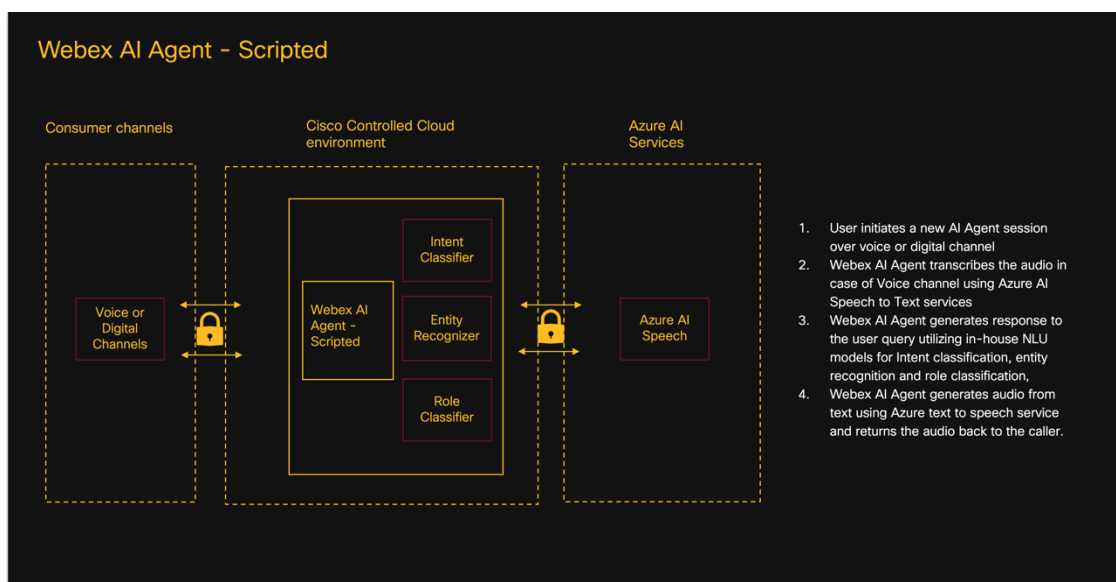
Communication with LLMs for data ingestion and responding to user queries is facilitated by Cisco's robust internal systems. These systems are designed to manage and optimize the allocation of LLM quotas, ensuring consistent availability and performance for customer interactions. Cisco employs advanced telemetry to monitor LLM performance metrics and regional availability in real time.

Cisco has established stringent data usage and privacy agreements with its LLM providers. These agreements explicitly prohibit the providers from utilizing customer data for their own model training or improvement purposes. This ensures that all customer data processed through the LLMs is not stored or retained by the providers for any purpose beyond the immediate transaction. PCI and PHI are redacted by default; clients can also choose to mask additional entities such as names, locations, health data, email addresses, phone numbers, and IP addresses.

The Interim Responses model leverages GPT 4.1-mini, Standalone Query Generator leverages a fine-tuned version of the facebook/bart-large-xsum, and the Turn Detection model is an ensemble model of the finetuned version of SmolLM2-135M-Instruct and bert-base-uncased, selected for its low-latency performance and ability to operate within strict inference time budgets. These models are hosted within Cisco's infrastructure and optimized for cost and responsiveness.

Scripted AI agents

Scripted AI agents are powered by conventional machine learning algorithms for natural language understanding. Users can choose between various NLU engines available on the platform – Swiftmatch and Mindmeld. In certain offerings, such as Webex AI Pro 2.0 (with Swiftmatch), a Large Language Model (LLM) is incorporated during training to enhance intent classification accuracy. Both engines are hosted in Webex Contact Center infrastructure to ensure that proprietary training data and models always remain within the customer's control. These NLU engines use the data configured by agent developers for training classifiers models.



Usage Guidelines

Autonomous and Scripted AI agents can be designed in Webex AI Agent platform and integrated with voice and digital channels using flow builders provided by Webex Contact Center and Webex Connect respectively. These agents are not intended for usage outside of these channels at the moment.

The model's output should be handled with strict confidentiality and ethical consideration, ensuring that data is solely used to improve customer satisfaction.

Outputs generated by AI and machine learning technologies may contain inaccuracies or errors. Depending on how they are used, these errors can lead to incorrect information or unintended actions. Users are strongly advised to validate critical outputs and take appropriate precautions when incorporating AI into their applications or decision-making processes.

Model Inputs and Outputs

Model inputs and outputs vary by agent type and purpose.

Autonomous AI agents

Model Inputs

- User query: A natural language question or statement that the agent needs to answer.

- Agent goal: The purpose of the AI agent defined by the user in natural language.
- Agent instructions: User defined instructions for the AI agent to complete its goal.
- Agent actions: The actions an AI agent can perform. These are defined by providing action description, type, and input entities required to complete the action.
- System instructions and guardrails: Additional instructions and guardrails provided by the platform to keep the conversation on track.
- Document corpus: A collection of documents (e.g., articles, reports, manuals, content from publicly accessible web URLs) that the AI agent can reference.
- Contextual information: Any additional context that might be relevant, such as previous interactions or user preferences.
- Standalone Query Generator: user query and conversation transcript.
- Interim Responses: System-level acknowledgment triggers, user query and call transcript.

Model Outputs

- Generated response: A comprehensive and informative response to the user's query, incorporating the extracted information or action performed.
- Standalone Query Generator: A generated standalone query which is used for further processing such as retrieval.
- Interim Responses: A contextual interim message to be played back to the user.

Scripted AI agents

Model Inputs

- User query: A natural language question or statement from the end-user.

Model Outputs

- Scripted Response: User configured Response depending on the intent classification and slot identified.

3. Data Sources for Training & Education

AI agents are trained on open-source datasets that are designed for information retrieval and question answering and do not use any Cisco customer data for training its models. Cisco does not use Customer Content to fine tune or train the models used to deliver AI agent responses. Microsoft represents that it does not use customer data to improve Azure OpenAI models and does not store or retain Cisco customer data in the Azure infrastructure. For details about the data source used by the underlying model from Microsoft's Azure OpenAI Service, please review the [Transparency note for Azure OpenAI Service](#) published by Microsoft. The Standalone Query Generator and Turn Detection models are trained using a combination of synthetic data generated via GPT-4.1 prompts and publicly available open-source datasets from platforms like Hugging Face. No Cisco customer data is used in the training or fine-tuning of these models. Scripted agents do not source data from any other source(s) than the one(s) provided by customers.

Data Locality for AI Agents Components

The following table shows the data locality of the various components that are a part of the AI Agents -

AI Agent Region	AWS	Azure STT/TT S	ElevenLabs	Deepgram	LLM Proxy (Internal)	Azure Open AI
produs1	N. Virginia, N. California	East US, West US	N. Virginia, N. California	N. Virginia, N. California	Ohio, US	US Regional + Global*
prodeu1	London, UK	UK South	EU regional	EU regional	London, UK	London, UK
prodeu2	Frankfurt, Germany	Germany West Central	EU regional	EU regional	Frankfurt, Germany	Frankfurt, Germany
prodca1	Canada Central	Canada Central	-	-	Canada Central	Global*
prodj1	Tokyo, Japan	Japan East	-	-	Tokyo, Japan	Global*
prodanz1	Sydney, Australia	Australia East	-	-	Sydney, Australia	Global*

prodsg1	Singapore	Southeast Asia	-	-	Singapore	Global*
prodin1	Asia Pacific (Mumbai), India	Central India	-	-	Sydney, Australia	Global*

*These requests can be processed in any data center, globally, where these models are hosted. More information about Azure Data Residency can be found [here](#).

4. Model Evaluation and Performance

Cisco frequently assesses and evaluates 3rd party models to maintain and/or improve both performance and accuracy of output. Humans are involved with the review, testing, and quality assurance associated with the deployment of the underlying models.

5. Safety, Accuracy, and Fairness

Safety and Accuracy

Cisco takes steps to mitigate safety risks and ensure ethical use when leveraging 3rd party models. Our vendors undergo rigorous reviews, which include security and safety assessments.

Please contact Cisco if you have concerns or feedback about your experience with Webex AI Agent.

Fairness

To understand more about how the 3rd party models are trained and evaluated for fairness, please refer to [Microsoft Azure Open AI Fairness note](#). Scripted agents are more deterministic in responses and detect only specific user intents based on the training data provided by customers themselves. This helps mitigate the risk that biased language is generated. Large Language Models (LLMs) can reflect societal biases present in the training data. These biases, related to factors like race, gender, and religion, can lead to unfair or discriminatory outputs from the AI agent.

6. Privacy and Security

Information about how we approach processing and security around personal data, including data retention period, etc., can be found in our [Offer Disclosures](#), found on the Cisco Trust Portal.

For voice interactions handled by AI Agents, customers may enable call recording.

As mentioned above, humans with the appropriate roles and permissions are involved in the review, testing, and quality assurance processes and may sample the inputs to and outputs from these models periodically.

7. Monitoring and Maintenance

Ongoing Cisco updates and product release documents include all significant product updates and any improvements in user experience. Cisco may also use customer email notifications and Webex Help Center documentation to inform customers of product changes.

8. References

- [Cisco RAI Framework](#)
- [Cisco Principles for Responsible AI](#)
- [Cisco Trust Portal](#)
- [User Guide](#)
- [Abuse Monitoring](#)
- [Content Filtering](#)

