



AI Technical Transparency Note

Template Completion Guide – 3rd Party or Hybrid Model

Owning Function: Webex Customer Experience Solutions

Owner(s)

Organization	Name (CEC ID), Title
Cisco	Ramya Tummala (ramyat), Product Manager

Revision History

Revision	Date	Name (CEC ID), Title	Comment
V2	1 sep 2026	Ramya Tummala (ramyat), Eng Product Manager	- Updated feature name from Suggested Responses to Real-time Assist - Updated model from 4o mini to 5.4 mini

Contents

Feature Overview	2
Model Overview	3
Model Architecture	3
Usage Guidelines	4
Model Inputs and Outputs	5
Data Sources for Training & Evaluation	6
Model Evaluation and Performance	6
Safety and Ethical Considerations	7
Fairness	7
Privacy and Security	7
Updates and Maintenance	7
References	7

Role / Title	Name (CEC ID)
Feature PM	Ramya Tummala (ramyat)
AI solution architects	Magdalena Patronska (mpatrons), Gourav Jain (goujain)

At Cisco, we believe that Artificial Intelligence (AI) can be leveraged to power an inclusive future for all. 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](#), which is based on six principles—Transparency, Fairness, Accountability, Privacy, Security and Reliability (the “Framework Principles”). Cisco translates the Framework Principles into product development requirements, which ultimately form part of the product development lifecycle alongside our Security by Design, Privacy by Design, and Human Rights by Design processes.

Accordingly, Webex by Cisco (Webex) features that leverage AI are built with transparency, fairness, accountability, privacy, security, and reliability at their core. Each feature powered by AI undergoes an AI Impact (AI) Assessment—a best-in-class review of how the technical underpinnings of the functionality measure against the Framework Principles.

The Real-time Assist feature is built with the Framework Principles at the center of how we deliver the AI-powered technology. This Technical Note describes more information about the feature, and how we responsibly leverage AI to deliver the functionality.

Feature Overview

The Real-time Assist feature is designed to assist human agents in contact center interactions by providing real-time, context-aware suggestions on how to respond to customer questions queries. This feature is displayed within the AI Assistant widget in the Webex Contact Center agent desktop.

Key functionalities include:

- **Contextual knowledge-based suggestions:** Provides suggested text for chat interactions and informative content for voice interactions, helping agents have informed conversations.
- **Action-based suggestions:** Suggests relevant actions that agents can either execute automatically (unmoderated mode) or review and approve (moderated mode).
- **Customizable AI Assistant Skills:** Administrators can create and configure AI Assistant skills in AI Studio, defining skill names, goals, instructions, knowledge bases, and actions. These skills can then be linked to specific queues.
- Real-time transcription for both voice and digital channels to generate timely and relevant suggestions.
- **Feedback Mechanism:** Agents can provide feedback (thumbs up/down) on the helpfulness of suggestions.
- **Auditability:** Provides administrators and developers with tools to audit and debug suggestions, including interaction transcripts, sources of suggestions, and action execution details.

The feature aims to improve agent efficiency, enhance customer experience, and ensure consistent service delivery by leveraging AI to provide immediate and relevant support to agents.

Model Overview

Model Architecture

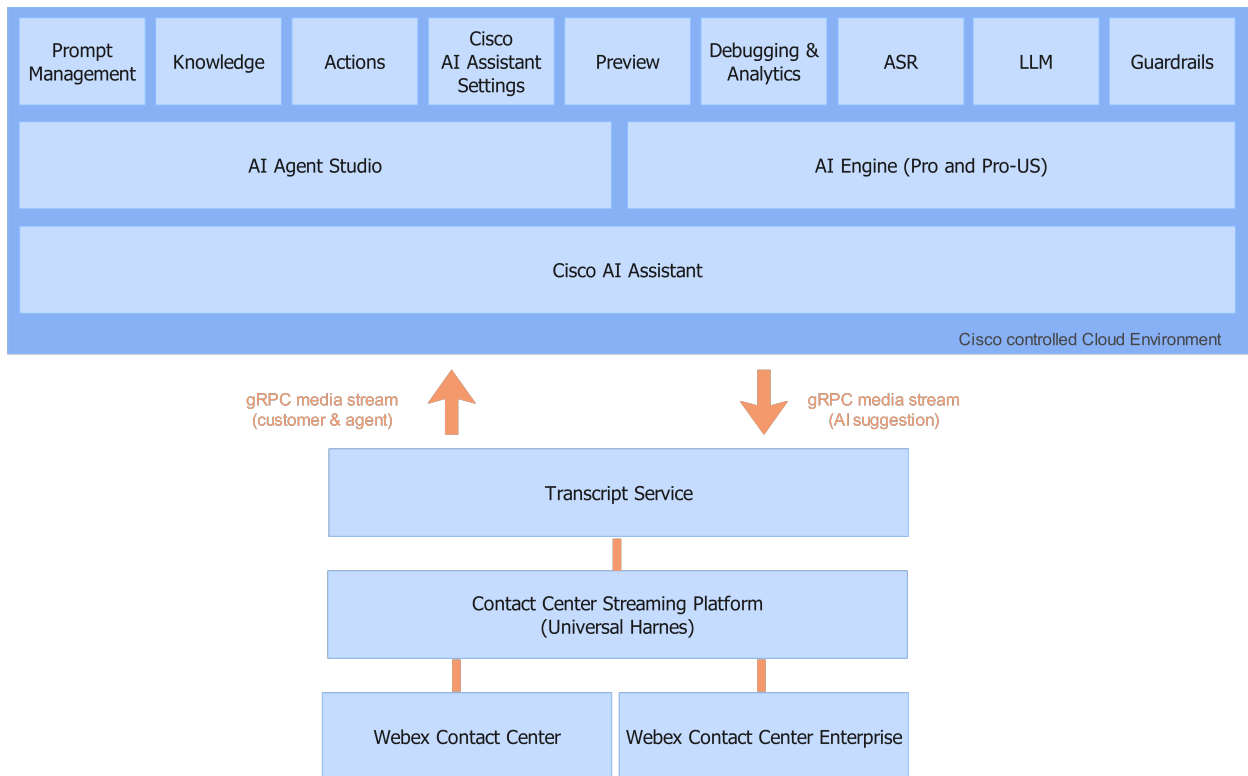
The Real-time Assist feature leverages AI agents for generating dynamic suggestions based on real-time interaction context.

It is powered by a third-party Large Language Model (LLM) from Microsoft's Azure OpenAI Service. For more details on Azure OpenAI Service and how it handles data, please review [Microsoft's transparency note available here](#).

The feature provides additional instructions, guardrails and context to the Azure OpenAI models to help ensure a well-orchestrated experience for the end users. The feature currently uses the 5.4 mini model offered by Azure OpenAI Service. Based on our research and benchmarking, we found this model to provide the best experience for the use case in terms of cost, latency, and quality. However, we are constantly evaluating available options; in the event of any material changes, we will update this document accordingly.

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 to help ensure consistent availability and performance for customer interactions. Cisco employs advanced telemetry to monitor LLM performance metrics and regional availability in real time. In the event of localized slowdowns or performance issues, our system dynamically reroutes customer requests to alternative regions that are both high-performing and compliant with relevant data protection standards, thereby maintaining seamless service continuity.

Furthermore, 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 helps ensure that all customer data processed through the LLMs is handled with the utmost confidentiality and is not stored or retained by the providers for any purpose beyond the immediate transaction.



Usage Guidelines

The Real-time Assist feature is designed to provide real-time guidance to agents handling customer interactions in voice and digital channels within the Webex Contact Center environment. The following guidelines apply to use of this feature:

- **Intended Use:** This feature is specifically designed for use within the Webex Contact Center agent desktop for active customer interactions and is not intended for use outside of these defined channels.
- **Unmoderated actions** are intended for low-risk tasks (i.e., actions where unintended fulfillment would not impact business operations or your customers). Examples include:
 - Retrieving or fetching relevant information;
 - Sending emails using pre-approved templates, where all fields have been reviewed.
- For actions that involve posting information to systems or modifying business records, we recommend configuring them as moderated actions. Moderated actions

require agent review of all fields and explicit approval prior to execution to ensure accuracy and to protect business operations.

- **Agent Oversight:** The AI-generated suggestions are intended as aids for the human agent. Agents are responsible for reviewing, validating, and, if necessary, modifying all Real-time Assist before using them. Critical outputs and actions, especially those impacting customer data or service, must be carefully verified by the agent.
- **Confidentiality and Ethics:** All information presented by the AI Assistant should be handled with strict confidentiality and ethical consideration and used solely to help improve customer satisfaction and interaction efficiency.
- **Potential for Inaccuracies or Errors:** Outputs generated by AI and machine learning technologies may contain inaccuracies or errors. Depending on how these technologies are used, such 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

- **Customer query:** A natural language question or statement from the customer, derived from real-time transcription of voice calls or digital chat messages.
- **Interaction transcript:** The ongoing conversation history, including customer utterances and agent responses, providing context for current queries.
- **Human agent response:** Agent's actions and adherence to/deviation from suggestions, which helps the AI adapt and refine future suggestions (e.g., providing a suggestion to follow a critical step if it was missed).
- **Skill goal:** The purpose of the AI assistant skill, defined by the administrator in AI Studio (e.g., "You are a helpful and polite assistant that will help human agents answer queries regarding lost baggage...").
- **Skill instructions:** Detailed step-by-step instructions configured by the administrator to achieve the skill goal, potentially referencing actions (e.g., "If the customer asks about lost baggage, ask for the PNR number and then <Retrieve_flight_details>...").
- **Actions:** Definitions of out-of-box or custom actions that the AI Assistant can perform, including Action Name, Description, Scope (Slot-filling only, Slot-filling and fulfilment), User Inputs (String, Number, Custom List, Regex, Date, Time, Phone, Email, Digit), and Fulfilment Flow.
- **System instructions and guardrails:** Additional instructions and guardrails provided by the platform to keep the conversation on track.
- **Document corpus (Knowledge Base):** One knowledge base selected by the administrator from a dropdown list of available knowledge bases, used by the AI agent for information retrieval.

- **Contextual information:** Additional context from the interaction, such as previous customer utterances or agent-provided additional context to refine suggestions.

Model Outputs

- **Generated response:** Suggested text for chat interactions or information for voice interactions, tailored to the customer's query. This response will be grounded in the knowledge bases and the skill instructions defined by the administrator. ~~and the AI Assistant's skill configuration.~~
- **Actions:** Recommendations for actions that can be executed automatically or require agent moderation, presented with relevant details and input fields.
- **Action execution status:** Agents receive confirmation when both moderated and unmoderated actions are fulfilled.
- **Collected information:** For moderated actions, information collected by the AI Assistant is presented in a review format to the agent.
- **Error details:** Information provided to the human agent if an action fails, allowing for correction and resubmission.

Data Sources for Training & Evaluation

The LLM 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 Real-time Assist. 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.

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.

Safety and Ethical Considerations

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](#). 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.

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.

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.

Updates and Maintenance

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

References

[General Disclosures](#)

- [RAI Principles](#)
- [RAI Framework](#)

- [Cisco Trust Portal](#)
- [Content Filtering](#)
- [User Guide](#)
- [Abuse monitoring](#)