Comstice

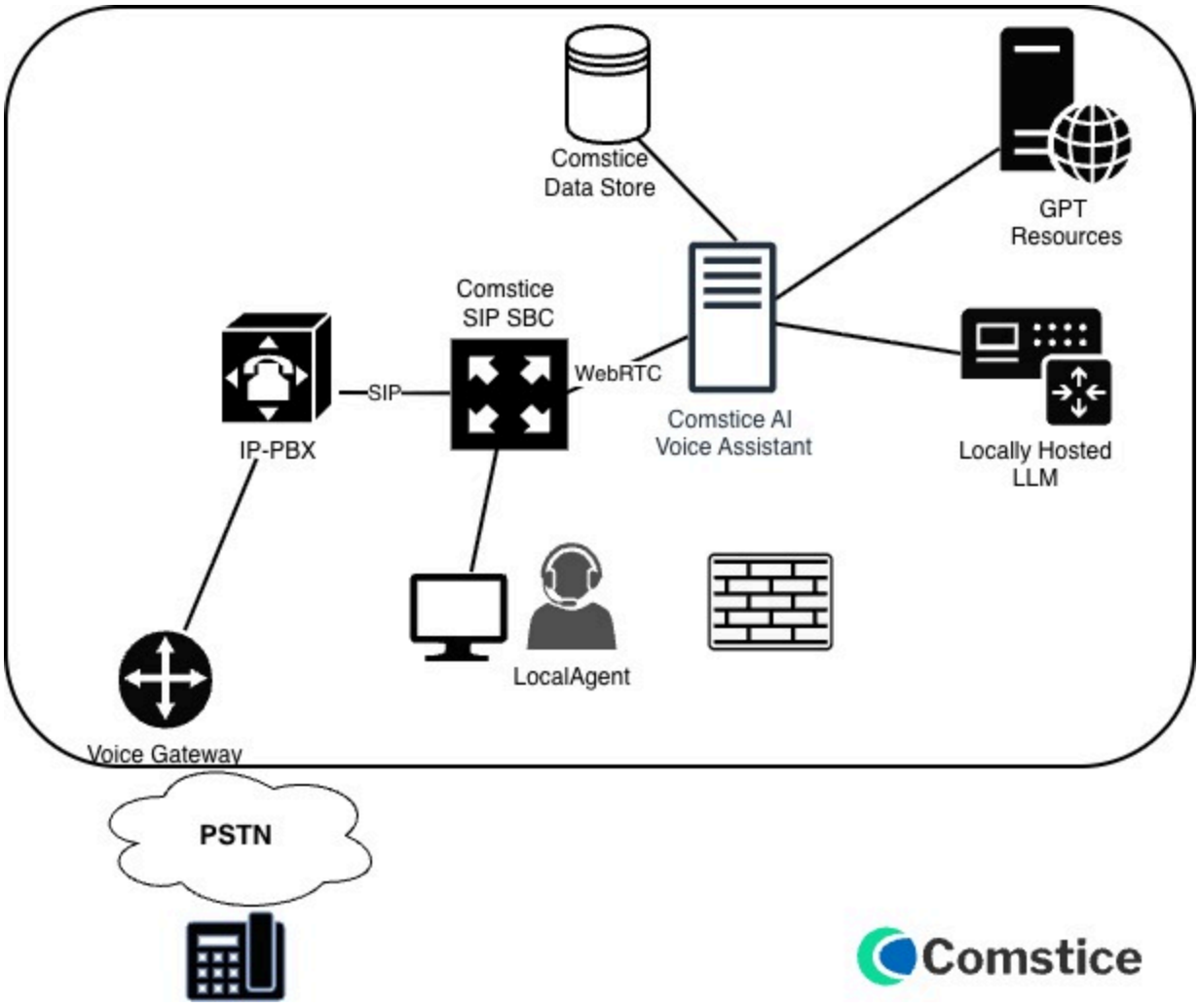
AI Voice Assistant for Cisco

Data Sheet

Comstice AI Voice Assistant

Comstice AI Voice Assistant includes;

- Speech to Text / Text to Speech Engine
- Vector Datastore
- Text Embedding Model
- Image Processing Model (For screenshots on the articles)
- Similarity Search Plugin
- Call Recording and call transcripts

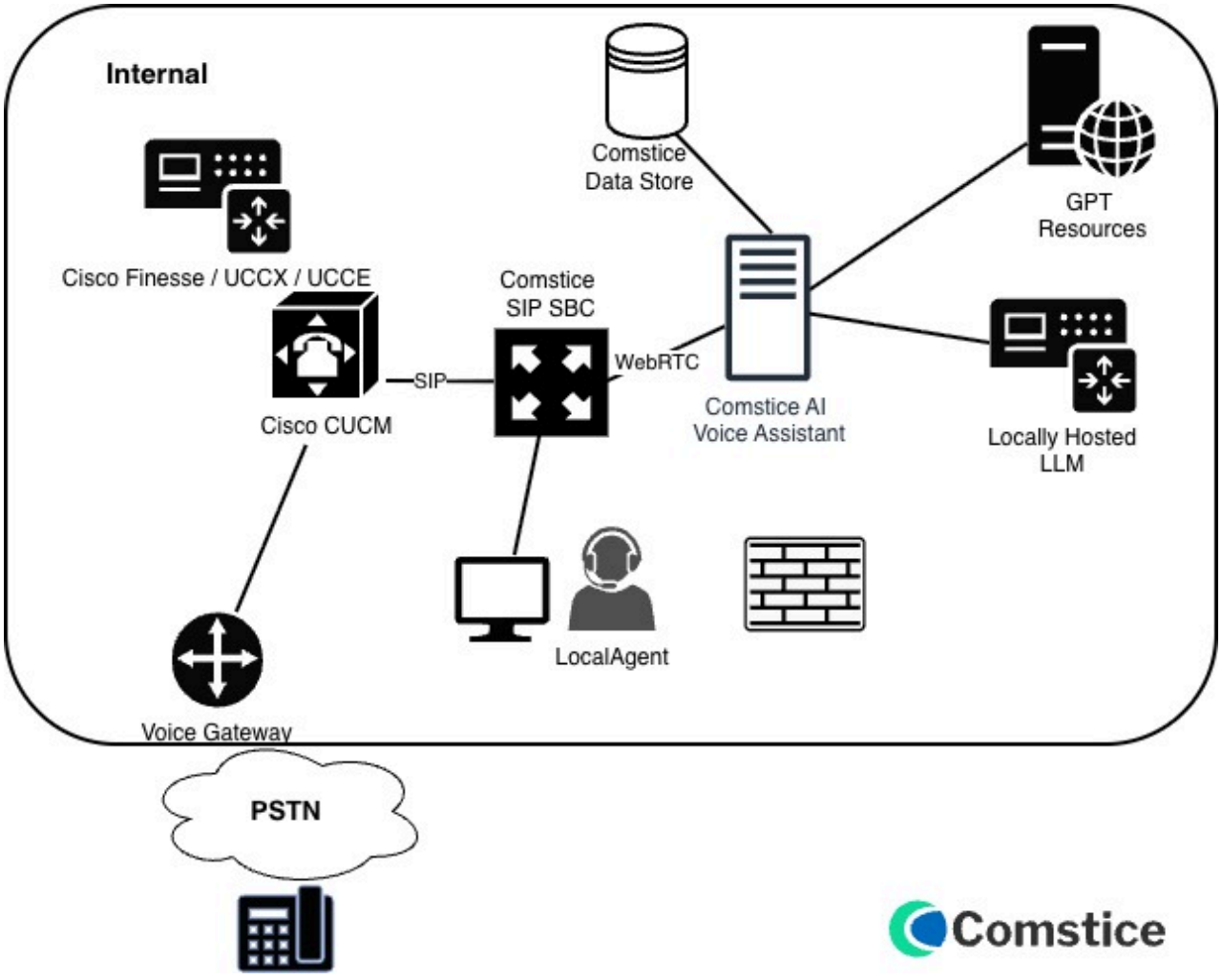


Comstice

Integration with Cisco

Comstice AI Voice Assistant will integrate with Cisco CUCM via SIP Trunk and handle calls for the created use cases.

If the call needs to be sent to a live agent, the call context can be added to the SIP Header as key/value pairs and then will be transferred using SIP REFER.



Three Layers of AI Voice Assistant

Comstice AI Voice Assistant has three layers in the use cases;

Layer 1: Simple IVR Replacement

User can pronounce the department and person names, then the call will be routed to the relevant extension or a queue.

This helps to avoid IVR tree updates, any confusions about IVR options and eliminates IVR recorded prompts.

Layer 2a: Prepared Questions and Answers

In this layer, administrators can enter the potential questions and answers into the system using the admin panel. If there is a match, the response is played. Also, you can ask whether the response was helpful or not. Negative responses will be marked and answers may need to be updated.

This helps to offer a certain level of self service without any complex LLM and GPU resources needed.

Layer 2b: Getting Answers from CRM, Database or API Integrations

You can also map the questions with the responses from third-party platforms such CRM services and databases.

Layer 3: Longer conversations with a context window

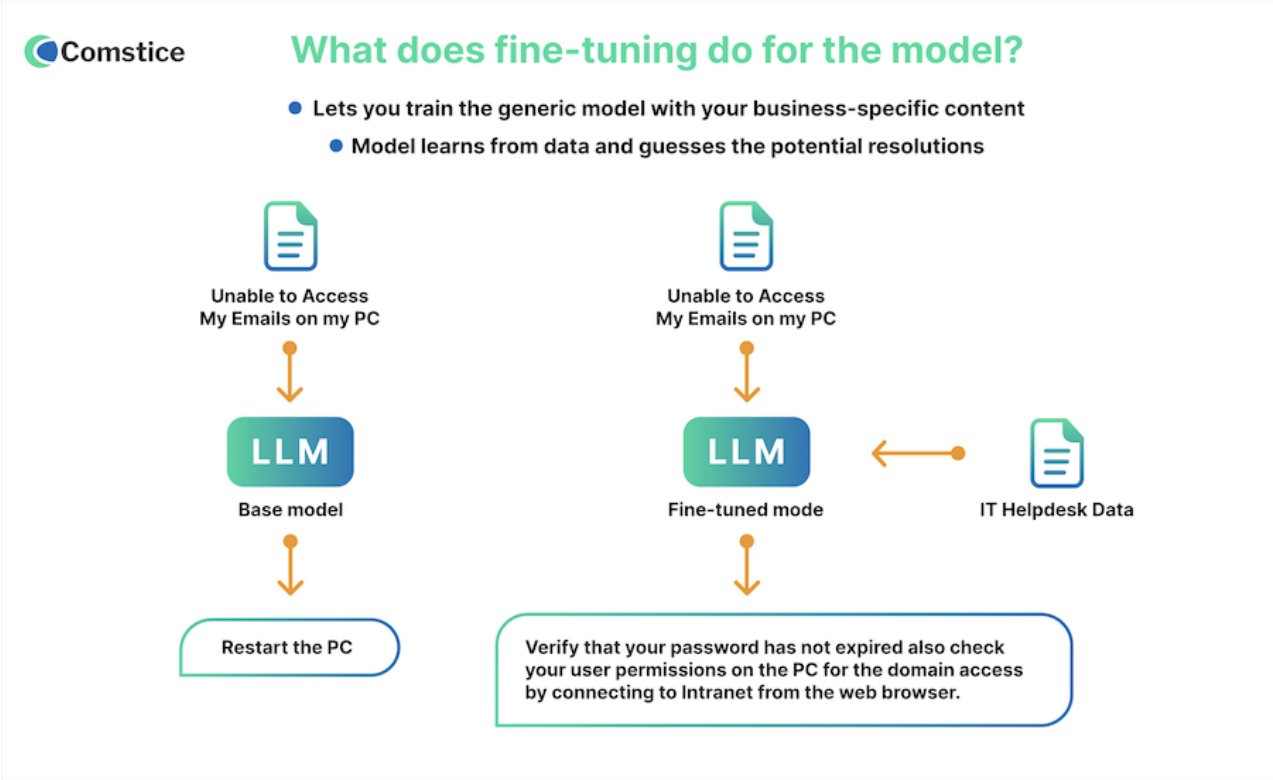
This is the layer you will need the recently created solutions for LLMs, Speech to Text (STT) and Text To Speech (TTS). Here, you have two options;

- **Cloud-Based LLM, STT, TTS Services:** They are often subscription-based but you need to expose your data
- **On-Premises LLM, SLM, STT, TTS Services:** In this option, you have the full data privacy but you will need GPU resources. For example, if you have 15,000 articles about IT Support, those articles need to be vectorized for **Fine Tuning** the LLM. LLM is nothing more than scraped public Internet content that is vectorized.

LLM Fine Tuning

Fine tuning is the process of training generic language models with your own content for your specific use cases. Knowledge base articles, blog posts, call transcripts of the recorded conversations with resolution can be used as the information stack.

After fine-tuning, the language model will be able to answer questions specific to your business.



Scope Your AI Voice Assistant

For Comstice to quote your AI Voice Assistant solution, you need to clarify the following points;

1. Which layer(s) are you planning to have? Layer 1 basic IVR, Layer 2 simple Q&A, Layer 3 conversational AI with context window and fine tuning
2. If Layer2b will be used, define the integration options for the third-party services and databases.
3. Whether cloud services should be used for LLM, STT, TTS or will they be all private for data privacy reasons?
4. If Layer 3 will be used what will be the size of the training data i.e. 10,000 articles with 2000 characters on each article on average?
5. If Layer 3 will be used, what will be the potential conversation length on the call ?
6. Concurrent number of AI Voice Assistant sessions
7. Languages to be supported

On-Premises AI Voice Assistant

Comstice AI Voice Assistant creates conversational AI use cases at scale and integrate with your customer service call flows. It helps different departments to create their own use cases separately and runs offline for maximum data privacy.

Use Cases:

Comstice AI Voice Assistant helps in the following areas;

- Questions to customer knowledge base, product helplines, self deployment assistance
- User guidance with question and answers and specific use cases
- Technical assistance, troubleshooting, diagnostics
- Updates on reservations, schedules, appointments
- Out of hours help on certain tasks
- Technical and engineering help on products and solutions

Features

Comstice AI Voice Assistant solution;

- can work standalone on your network, with speech recognition and text to speech for multiple languages and dialects.
- with **built-in call recording**, we can analyze and optimize the performance of the natural voice conversations.
- can integrate with the AI agents created by Comstice for your use cases
- an integrate with LiveKit AI agents created by your team using LiveKit Agent Builder
- integrates with any SIP-Supported telephony service such as Avaya Session Manager, Cisco CUCM as well as the cloud solutions such as Five9, Nice InContact and, Genesys Cloud.
- Using the questions and queries collected from AI Voice Assistant, we can find out about the new features, common issues, features that are confusing to the user and improve our products and solutions.

AI Voice Assistant as a Service

Comstice AI Voice Assistant is also available as a service in the cloud. It can integrate with your on-premises or cloud telephony and provides the framework for the AI self service use cases

On-Site LLM vs Cloud LLM:

You can use on-site LLM resources if your use cases are at the Layer 3 of the AI Voice Assistant scenarios where the conversations are long and the context must be preserved during the call.

We can use locally-hosted GPU resources or the services such as Amazon Bedrock and RunPod GPU services.

Speech to Text (STT) and Text To Speech (TTS)

Comstice offers on-premises STT and TTS services that you can run fully offline. Depending on the languages and accents, we can also hook into STT and TTS services which may offer a wider variety of languages and regional dialects

Training TTS for a language or a dialect

If there is a particular language or a dialect you want to train, Comstice can help with that as well. We provide written text that a talent can read and record in the preferred languange. A TTS model can be trained by using this recorded audio.

Use Case: Convert Technical Articles into AI Voice Assistant

Comstice can help you to create your technical articles to AI Voice Assistant using locally-hosted solution;

Use Cases:

- ask questions about the products, parts, and features
- troubleshoot incidents
- check and order parts on the fly
- verify technical roadmaps, compatibility and feature information
- get suggestions about the quick fix and long-term resolution options

Process:

Comstice will do the following;

1. convert 1000+ articles into searchable knowledge base
2. Comstice Retrieval Tool
3. Integrate Retrieval Tool with the language model
4. Tune the system for voice: re-train the response for voice conversations. AI Chat Bot responses are often not suitable for a voice chat; they include bullet points markdowns etc. Comstice helps to convert that into a natural voice conversation;
5. short, spoken-sounding answers
6. asking clarifying questions instead of guessing
7. reading part numbers, terms, IDs slowly and phonetically
8. for longer responses, offer to send an article, paragraph or checklist by email
9. Improve honesty in the language model and avoid guessing when there is no viable response
10. Speech to text and text to speech optimization for technical terms and jargon
11. Manage the Latency Budget: Targeting under 800ms from the end of user speech to start of the agent speech.

Use Case: Stock Checks, Parts Ordering with AI Voice Assistant

Comstice can help you to query customer questions inside your stocks and orders database and AI Voice Assistant using lcoally-hosted solution;

Use Cases:

- verify the calling customer, PIN number, identity
- check stock information for the requested part
- verify estimated shipping cost and timelines
- handle purchasing and send automated verification emails

Process:

Comstice will do the following;

1. analyze the database structure and create a list of potential questions and SQL query answers
2. summarize the database response
3. cache the data where needed for faster response
4. log every tool call with inputs outputs caller ID and the call transcript
5. timeout and delay phrases such as "let me check on that" and keyboard noise, instead of a silent wait on the line
6. rate-limit the writes per call to avoid any security issues and attacks
7. retain the database connection throughout the service duration
8. for longer responses, offer to send an article, paragraph or checklist by email
9. Improve honesty in the language model and avoid guessing when there is no viable response
10. Speech to text and text to speech optimization for technical terms and jargon

AI Voice Assistant FAQ

Can I use this solution with cloud contact centers such as Five9 and Genesys Cloud?

Yes. Comstice AI Voice Assistant can be integrated with any SIP-supported service

Do I need to purchase GPUs? Can we just use the services available?

You don't need to purchase GPUs. First, you should check whether your use cases are similar to Layer 3 cases where context of the conversation must be preserved. Then services like Amazon Bedrock or RunPod can be used.

What is the difference between using local STT and TTS vs. STT and TTS as a service?

Comstice offers local STT and TTS for popular languages such as English, Spanish, Arabic, French, German, Italian etc.

If you are looking for a language or a dialect that is less common, STT and TTS services can be used.

For TTS, it is also possible to train your own model using a talent. Comstice can provide reading material that the talent can read and record. It is possible to train a TTS model using 1-5 hours of recorded audio, depending on the complexity of the conversations.

What is the license model?

License model depends on various factors such as where will the solution be hosted, whether it will be on-premises, whether any cloud LLM and STT/TTS services be used etc.

Visit https://comstice.com/request-callback to schedule a discussion with our experts.

Support

Comstice offers a break-fix support for all the solutions. Priority support is included in all the subscriptions as long as the solution is deployed with N+1 redundancy. SLA is one hour response and up to four hours fix, based around all the depedencies of each solution. Support is available 24/7. Tickets can be raised by the client's IT personnel that has already followed the troubleshooting steps provided during the Administrator Training delivered as part of the project. Comstice does not accept tickets directly from the end users.

Tickets can be opened from https://comstice.com/support, by sending an email to support@comstice.com or by calling +1 713 929 3714 (Option 2)

Reporting questions, configuration and design questions are not part of the SLA and will be handled during regular office hours. Only break-fix support is handled on the SLA with 24/7 coverage.

Comstice

Thank You

Please contact sales@comstice.com
for demos and callback.