CES: Speech (STT, TTS, Voices) & Agent Framework

March 27th, 2025

Google Cloud | American Express | Five9

Pre-Workshop Sessions (virtual)

These sessions will be recorded and shared with the team to work around everyone's calendars.

An Introduction to core Five9 CCaaS Components (March 21st) [60 mins]

- Contact Routing
- WEM
- Outbound Dialer
- Agent Workspace
- Call recording

CES: Speech (STT, TTS, Voices) and GenAI foundations initiatives (March 27th) [60 mins]

- Review key customer and agent facing AI features

CES + Five9 feature and technical Overview (April 1st) [60 mins]

Agenda Thursday March 27th (EDT) - Remote

	Time	Session Title	Duration	Overview
1	10:55 - 11:00		5 min	Arrivals
	11:00 - 11:05	**Kick-off & Google Welcome**	5 min	Google | Lily Britt - Account Executive Google | Gloria Vargas - AI Specialist
2	11:05 - 11:40	• Speech to Text • Text to Speech • Translation • Conversational AI Agents	25 min	Google | Tomás Coyne – Head of Applied AI Solutions Google | Catherine Lori – Google Speech Specialist
3	11:40 - 11:50	• Gen AI: Agents	10 min	Google | Gloria Vargas - AI Specialist
4	11:50 - 12:00	Google & Five9 Session 3 Snapshot	5 min	Google | Lily Britt - Account Executive Five9 | Derek Bakker - Sr Majors Sales Director

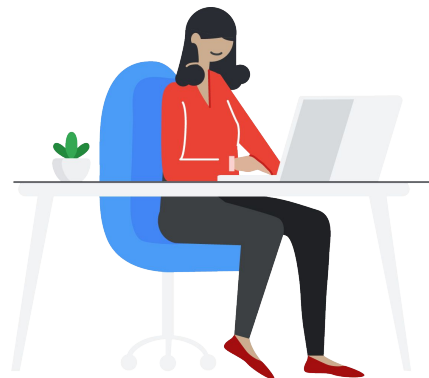AMERICAN EXPRESS | Five9 | Google Cloud

Text-to-Speech:

With industry-leading voice quality, Google Cloud's **Text-to-speech**

enables a natural, conversational manner across channels.
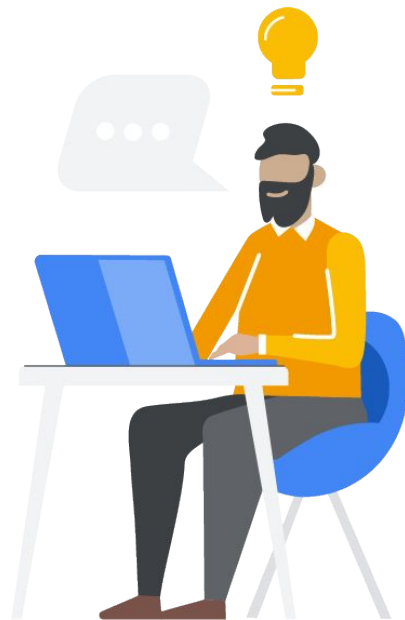
NEW OOTB Chirp 3: HD Voices

Represents Range of Pitch and Texture

Name	Sample
Aoede	🔊
Orus	🔊
Charon	🔊
Leda	🔊
Fenrir	🔊
Kore	🔊
Puck	🔊

Chirp 3: Instant Custom Voice

Trained with just 3-10 seconds of audio...

	Customer Service (Catherine)	Sports Broadcaster	Tomás Voice
Training Audio:	🔊	🔊	🔊
Cloned Sample:	🔊	🔊	🔊 🔊

Speech-to-Text

With unparalleled accuracy and real-time processing, Google Cloud's **Speech-to-Text** transforms audio into actionable insights across platforms.
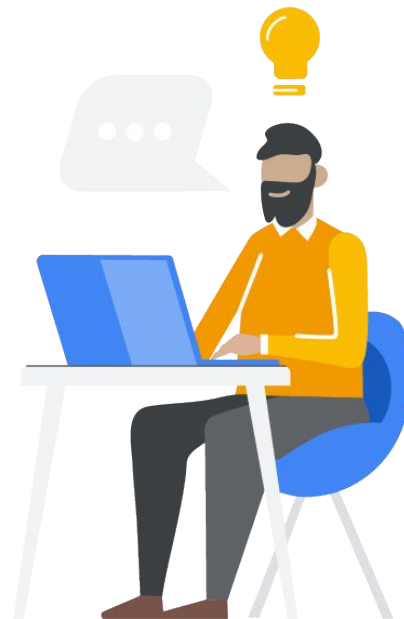
Speech-to-Text API

Audio Transcription

Industry leading accuracy
Highly accurate and robust to audio types and quality

Support for 300+ languages in 140+ variants
Transcribe speech from anywhere or auto-detect the language

Model Adaptation

Customize results for domain

Automatic Punctuation

Spoken our automatic

Diarization

Separate speakers automatically

Domain Specific

Models for telephony and medical

Reduction in Word Error Rate (WER) from 25% to <4%, as compared with legacy STT models
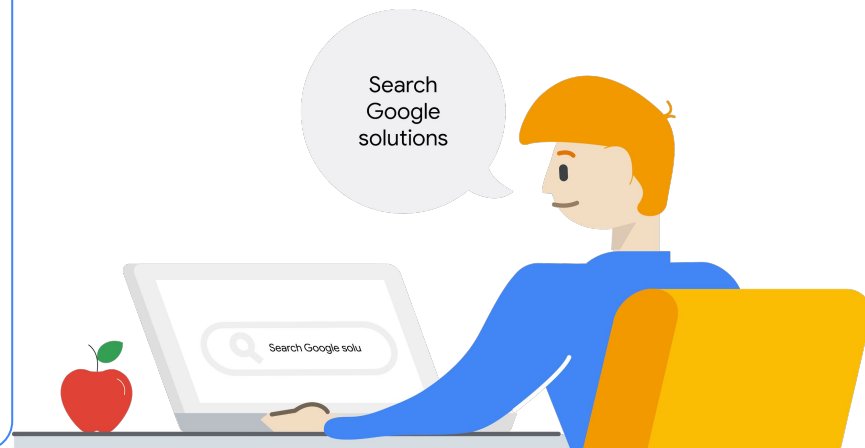
A brief history of speech at Google

RNN-T

MultiDomain

First Deep
Neural Network

First
LSTM
Network

First Sequence
to Sequence
Network

Transformer

First Google
speech patent

Voice search in
40+ languages

Conformer *

| 2001 | ...... | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2019 | 2020 | 2022 |
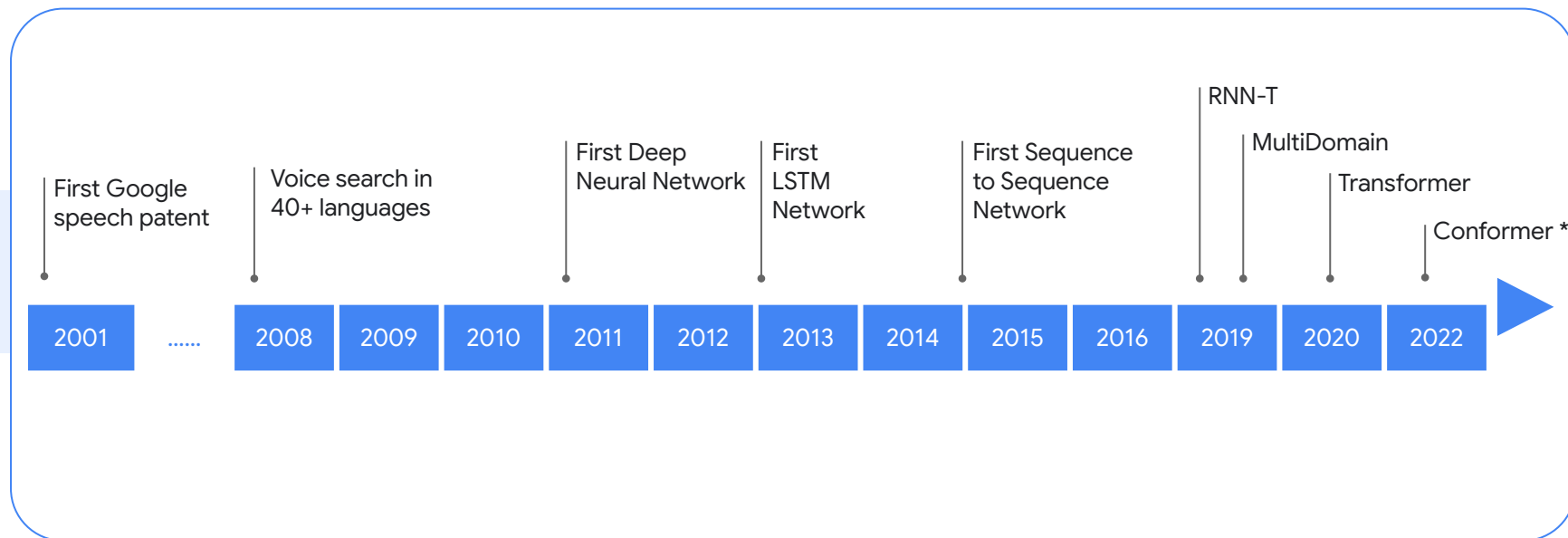
Today: Chirp for Speech-to-text

Build voice-enabled applications for global audiences

Apply the power of large models to Speech
Self supervised training on millions of hours of audio to create the first 2B parameter speech model.

Improve speech recognition quality
Achieves 98% accuracy on English and over 300% relative improvement in tail languages. Bringing the quality of all the world's languages closer to that of widely spoken languages.

Unlock speech use cases and languages
From captioning of Punjabi videos to Balinese voice assistance, users can now build conversational experiences in any language.

AMERICAN EXPRESS | Five9 | Google Cloud
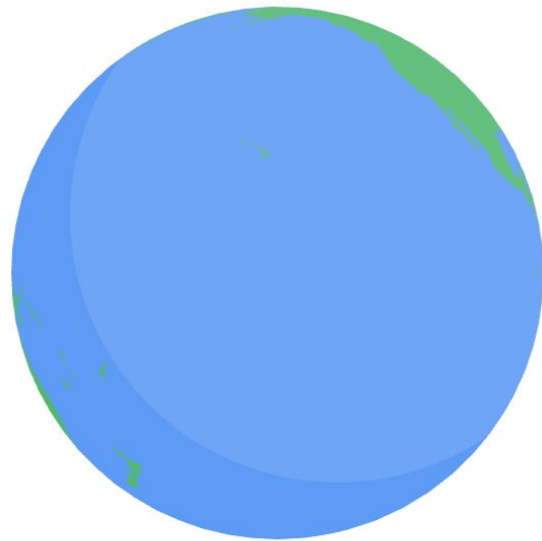
GenAI - AI Agents

Why agents on Google Cloud

01

Higher quality

Achieve higher performance for real world utility through our vertical integration across client, server, and model

02

Production ready

Build agents with the platform we stand behind and use to build our own solutions.

03

Built for enterprise

Handle enterprise requirements with HiTL flows and guardrails, policy enforcement, and Google Cloud security & compliance

04

Open Ecosystem

Enjoy interoperability to future proof investment and bridge the gap to existing popular frameworks tools and connectors

05

One stop shop

Accelerate development with all you need to build agents: frameworks, tools, runtimes, active learning, solutions and models (Gemini, Vertex AI Model Garden, fine tuned models)

Simplify the process of building and deploying agents

Helping developers discover, build, and deploy agents

Discover, learn, and get started building agents using samples and tools in the Vertex AI **Agent Garden**

Define, test, and rapidly develop multi-agent applications using Google's open source **Agent Framework**

Deploy, manage, and scale agents using the Vertex AI **Agent Engine**

Evaluate, monitor, trace your agent's behavior for continuous improvement and auditing

Discover, build and deploy agents

Agent Framework
Client side SDK to define multi-agent applications for complex, real world scenarios

| Tool use | Orchestration |

Agent Engine (Fully-managed runtime)

| Monitoring and Logging | Example Store | Session Management |
| Evaluation | Long-term memory | Agentic Autorater |

Agent Garden
Templates and tools to accelerate agent development

Quality begins with models optimized for agents

Gemini excels in agentic performance because:

Model Capabilities

multimodality, long context, and minimal hallucinations

Advanced Reasoning & Planning

Core training focus, crucial for agentic quality

Tool & Code Integration

Advanced function calling, including compositional function calling

Factual Grounding

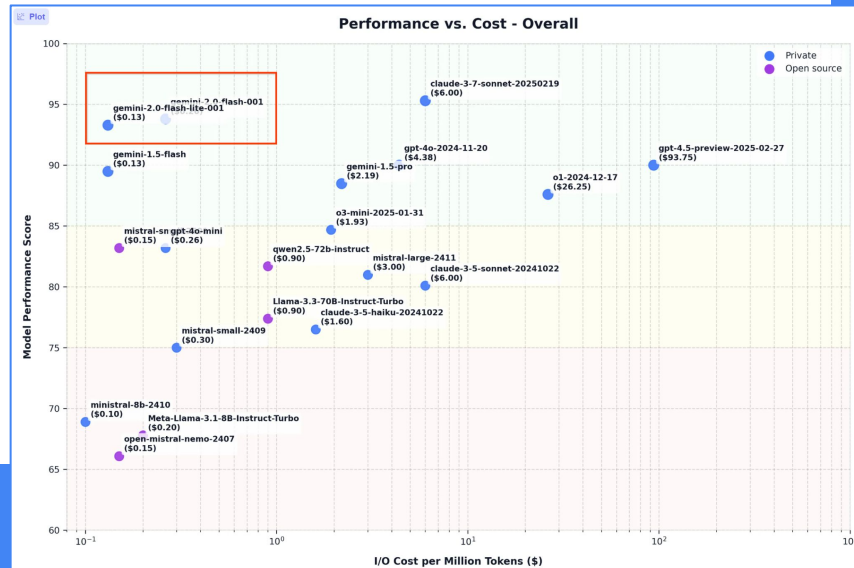Leading factual accuracy and programmatic instruction following

Effective Evaluation Methodology

End-to-end testing mimicking real-world agent flows

Agent Leaderboard

GenAI is evolving rapidly, with developers building exciting, high ROI agents. We built this leaderboard to answer one simple question:

"How do top LLMs perform in real-world agentic scenarios?"

Performance vs. Cost - Overall

- Private
- Open source

Model Performance Score (y-axis, 60–100)
I/O Cost per Million Tokens ($) (x-axis, 10^{-1} to 10^3)

- gemini-2.0-flash-lite-001 ($0.13)
- gemini-2.0-flash-001
- claude-3-7-sonnet-20250219 ($6.00)
- gemini-1.5-flash ($0.13)
- gpt-4o-2024-11-20 ($4.38)
- gpt-4.5-preview-2025-02-27 ($93.75)
- gemini-1.5-pro ($2.19)
- o1-2024-12-17 ($26.25)
- o3-mini-2025-01-31 ($1.93)
- mistral-sn ($0.15)
- gpt-4o-mini ($0.26)
- qwen2.5-72b-instruct ($0.90)
- mistral-large-2411 ($3.00)
- claude-3-5-sonnet-20241022 ($6.00)
- Llama-3.3-70B-Instruct-Turbo ($0.90)
- claude-3-5-haiku-20241022 ($1.60)
- mistral-small-2409 ($0.30)
- ministral-8b-2410 ($0.10)
- Meta-Llama-3.1-8B-Instruct-Turbo ($0.20)
- open-mistral-nemo-2407 ($0.15)

AMERICAN EXPRESS | Five9 | Google Cloud

Agent Framework

Client side code framework and SDK to build sophisticated, multi-agent solutions

Define **multi-agent** applications orchestrating actions across many agents and tools

Interleave **deterministic l**ogic with actions driven by **gen AI reasoning** for effortless, hybrid agents

Intuitive local **dev UI** for fast iteration; visualize agent topology and trace agent's actions
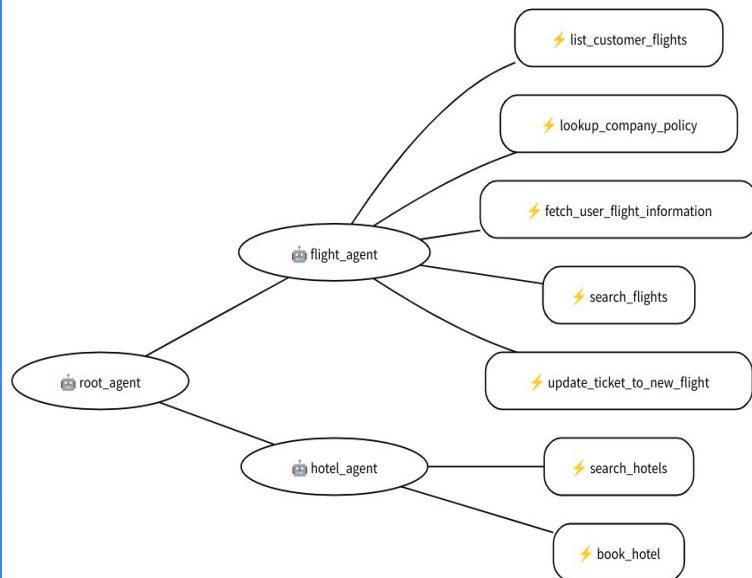
Gemini is the default, but **any generative AI model** is supported, including fine-tuned models

Live audio, video, screen bi-directional streaming

Build your agents, how we build agents.

Agent Definitions

Graph   root_agent   flight_agent   hotel_agent

- 🤖 flight_agent
  - ⚡ list_customer_flights
  - ⚡ lookup_company_policy
  - ⚡ fetch_user_flight_information
  - ⚡ search_flights
  - ⚡ update_ticket_to_new_flight
- 🤖 root_agent
- 🤖 hotel_agent
  - ⚡ search_hotels
  - ⚡ book_hotel

Connect agents to tools in the agent framework

Quickly access tools, APIs, applications and databases via APIs to perform real-world tasks

- Out-of-the-box support for
  - Google Search
  - Code Execution
  - Vertex AI Search
  - Vertex AI RAG Engine
  - Big Query
  - GCP Databases
- Support for custom Functions and APIs
- Support for "Async Tools" for long running operations

Coming soon: API Hub integration

- **Managed connections** to popular enterprise applications, services and databases
- Connect with your custom APIs within Apigee

Coming Soon in April

Active Directory	Amazon S3	Apache Cassandra
Apache Kafka	Apache CouchDB	AuthorizeNet
Azure Data Lake	BigCommerce	Box
CockroachDB	Commercetools	Couchbase
Dynamics 365	Elasticsearch	Email
EnterpriseDB	FinancialForce	FreshBooks
FTP	Greenplum	HTTP
HubSpot	IBM MQ	Jira Cloud
Kintone	LDAP	Magento
MailChimp	MariaDB	MetaAds
MongoDB	MySQL	Netsuite
Neo4J	Oracle DB	PayPal
PostgreSQL	Rabbit MQ	Redis
Redshift	Salesforce	Salesforce Marketing Cloud
SAP ERP	SAP Netweaver Gateway	SAP HANA
SAP SuccessFactors	SendGrid	ServiceNow
SFTP	SharePoint	Shopify
SingleStore	Snowflake	Splunk
SQL Server	Streak	Stripe
TaxJar	Teradata	Tibco
Trello	Twilio	Webhook
Workday	Xero	
Zoho Books	Zuora	

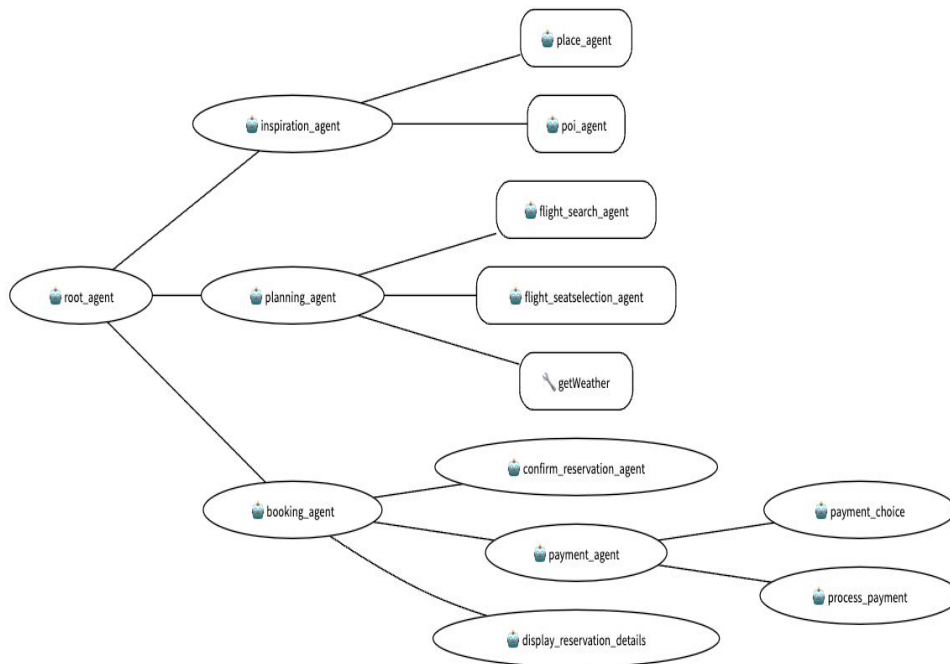AMERICAN EXPRESS | Five9 | Google Cloud

Travel Agent

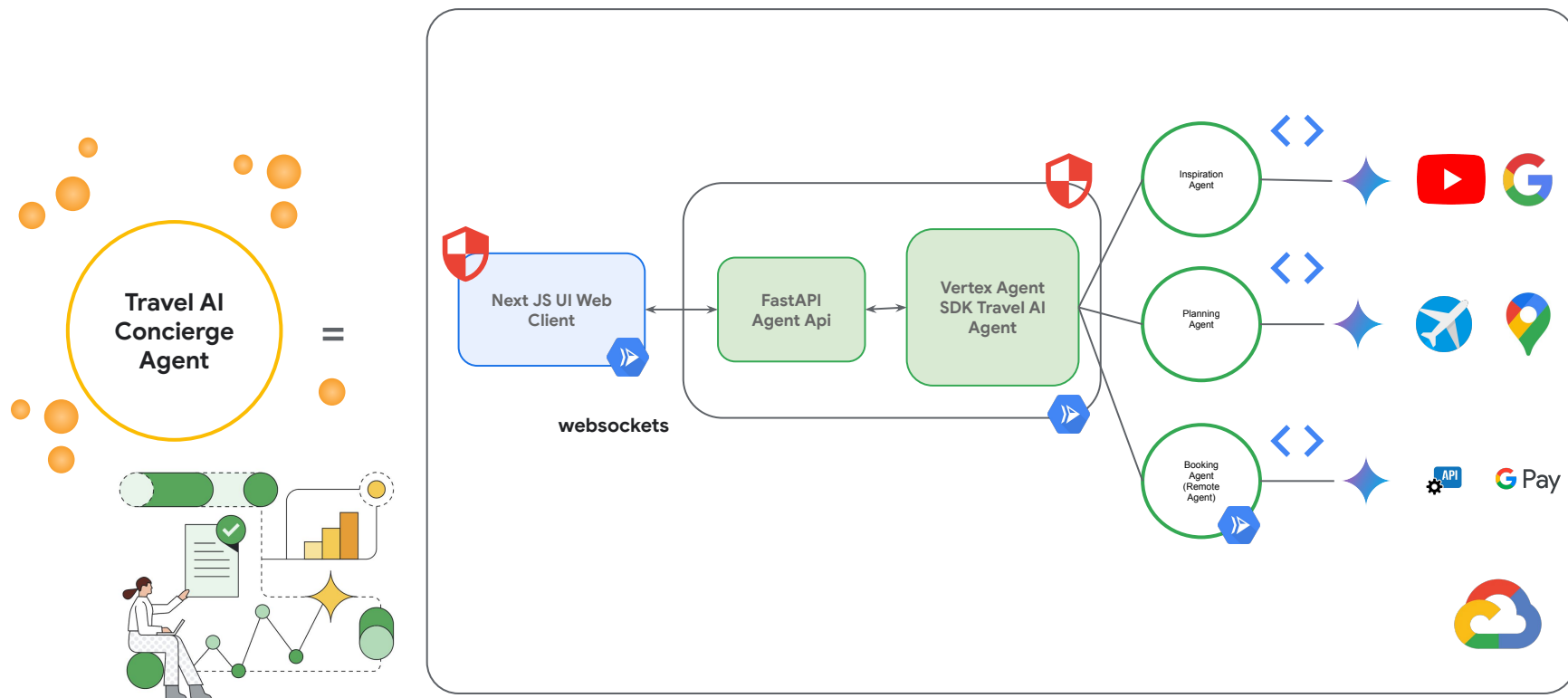# Travel Concierge Demo
Powered by the Agent Framework
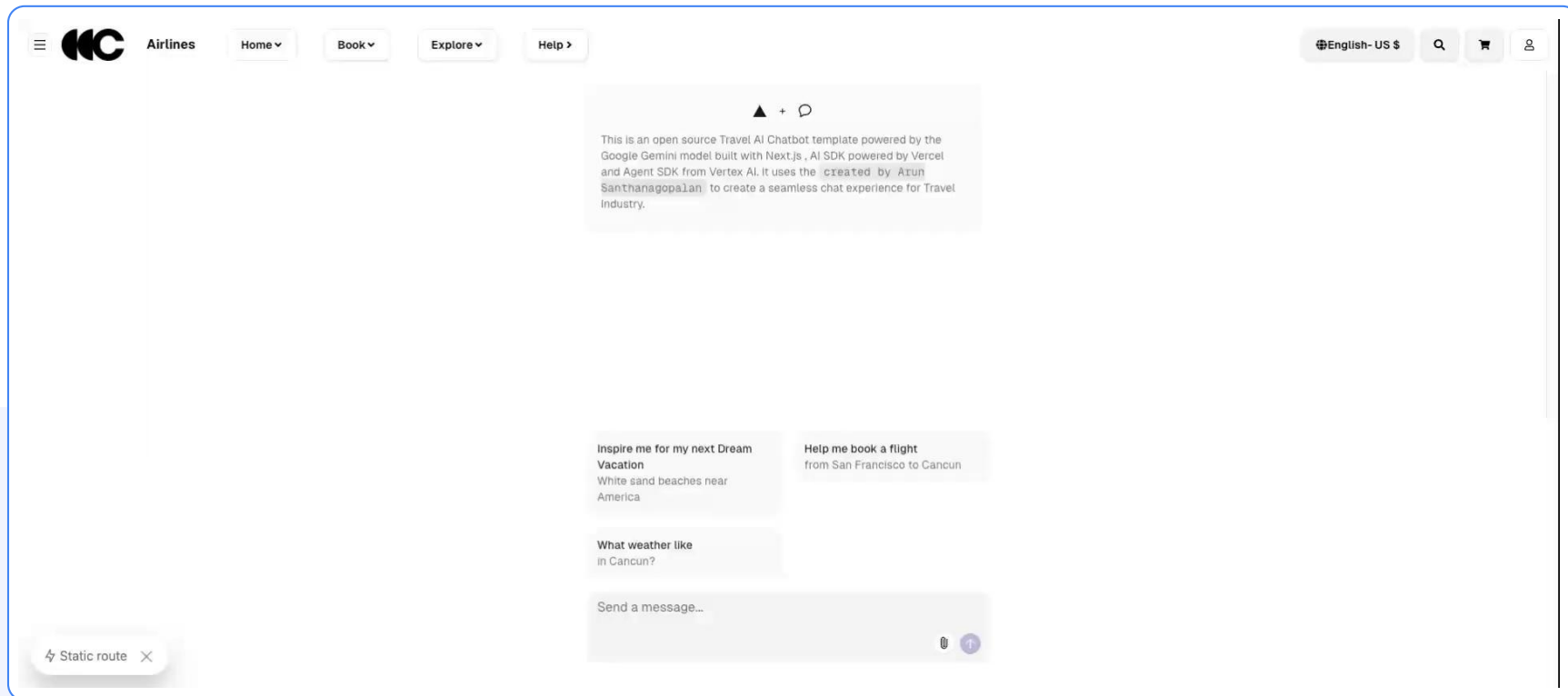
Agent Definitions

Graph root_agent inspiration_agent planning_agent booking_agent confirm_reservation_agent payment_agent payment_choice process_payment display_reservation_details

Agent Architecture

Pre-Workshop Sessions (virtual)

These sessions will be recorded and shared with the team to work around everyone's calendars.

An Introduction to core Five9 CCaaS Components (March 21st) [60 mins]

- Contact Routing
- WEM
- Outbound Dialer
- Agent Workspace
- Call recording

CES: Speech (STT, TTS, Voices) and GenAI foundations initiatives (March 27th) [60 mins]

- Review key customer and agent facing AI features

CES + Five9 feature and technical Overview (April 1st) [60 mins]

Thank You