

Technical atlas

Architecture, contracts, execution and deployment.

A production foundation. A clear direction.

CAI connects existing telephony to governed AI conversations, interchangeable speech intelligence and human supervision. The end-to-end voice path is in production.

Integrated queues, a live collaborative desktop and supervised infrastructure control extend this foundation through the platform roadmap.

In this edition

- 01 The platform
- 02 Responsibilities & planes
- 03 Governed execution
- 04 Speech & WebRTC
- 05 Tools & sensitive data
- 06 Connection choices
- 07 Outbound execution
- 08 Fleet improvement
- 09 Data ownership
- 10 Deployment & resilience
- 11 People & operating loop
- 12 Product workspaces
- 13 Roadmap & full CAI operation
- 14 Capabilities & terminology

The platform

CAI is a production platform for customer experience and contact-center integration. It connects existing telephony to governed AI conversations, interchangeable speech intelligence and a continuous cycle of training and improvement.

Five core components work together: **CAI Application & Control Plane**, **CAI Agentic Runtime**, **CAI Speech Platform**, **CAI Telephony Edge** and **CAI SIP Media Bridge**. In-house WebRTC brings the same speech capabilities to browsers. **CAI Adversarial Caller** exercises journeys through challenging conversations.

Organizations can add CAI to an existing contact center while retaining their queues, desktops and enterprise systems. The roadmap extends this foundation into a complete CAI operating environment with integrated work distribution and a collaborative agent desktop. CAI Speech Platform can also serve applications independently.

Core components

Component	Responsibility	Value
CAI Application & Control Plane	Organizations, journeys, training, business records and fleet decisions	One business and user root across the operation
CAI Agentic Runtime	Live journey state, response control, tools and execution workers	Business policy governs every consequential action
CAI Speech Platform	Realtime WebSocket, WebRTC, speech profiles and engines	Model choice through a consistent voice interface
CAI Telephony Edge	FreeSWITCH signaling, media anchoring and call control	Programmable integration with carriers, PBXs and contact centers
CAI SIP Media Bridge	SIP/RTP adaptation to TwiML-compatible media events	Direct telephony connectivity and provider flexibility

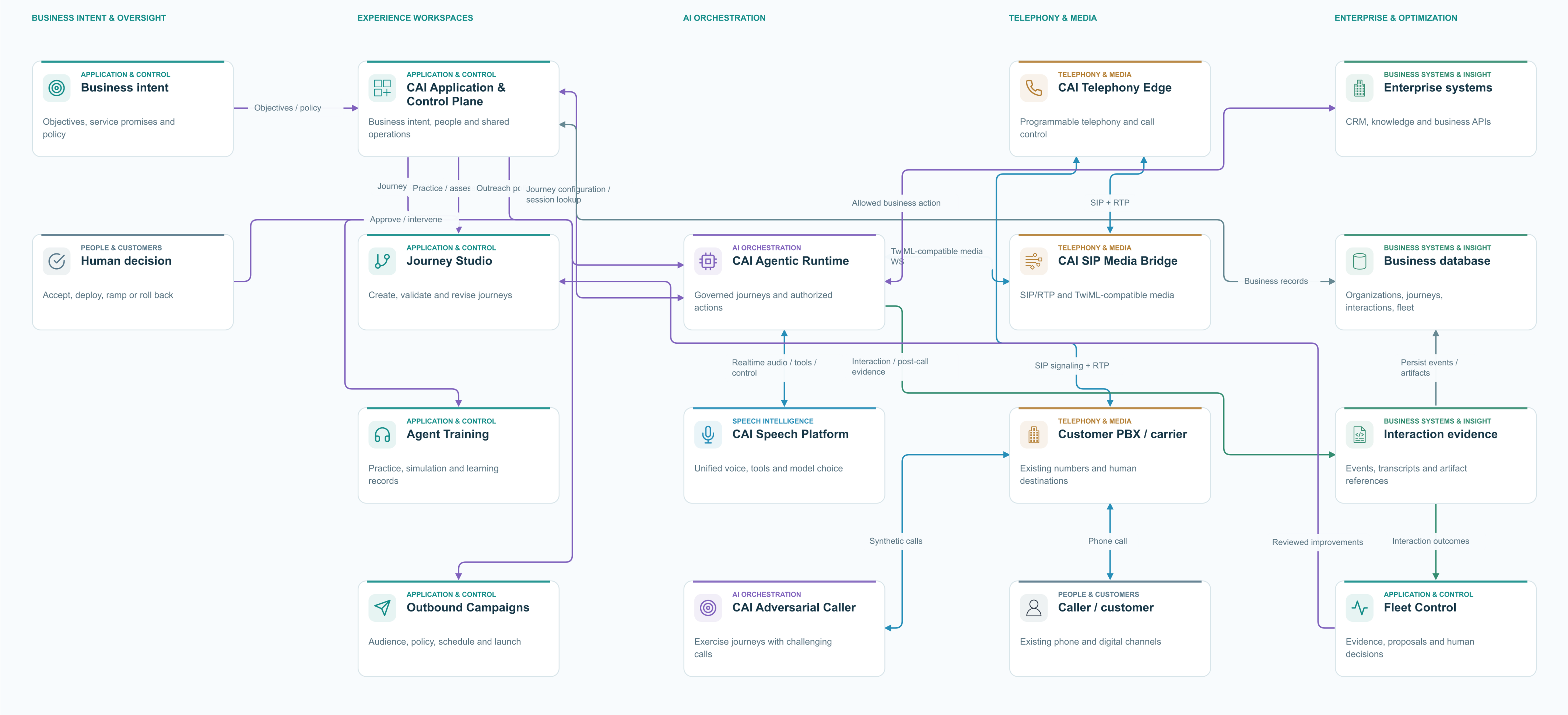
Production foundation and expansion

The end-to-end PBX, SIP bridge, runtime and speech path is in production. Journey execution, speech configuration, outbound processing and fleet improvement provide the foundation for connected AI operations.

The next stage adds durable customer work queues, live agent assignment and supervised infrastructure changes. These extensions reuse the same journeys, speech services, business records and human control.

D01 · Platform topology

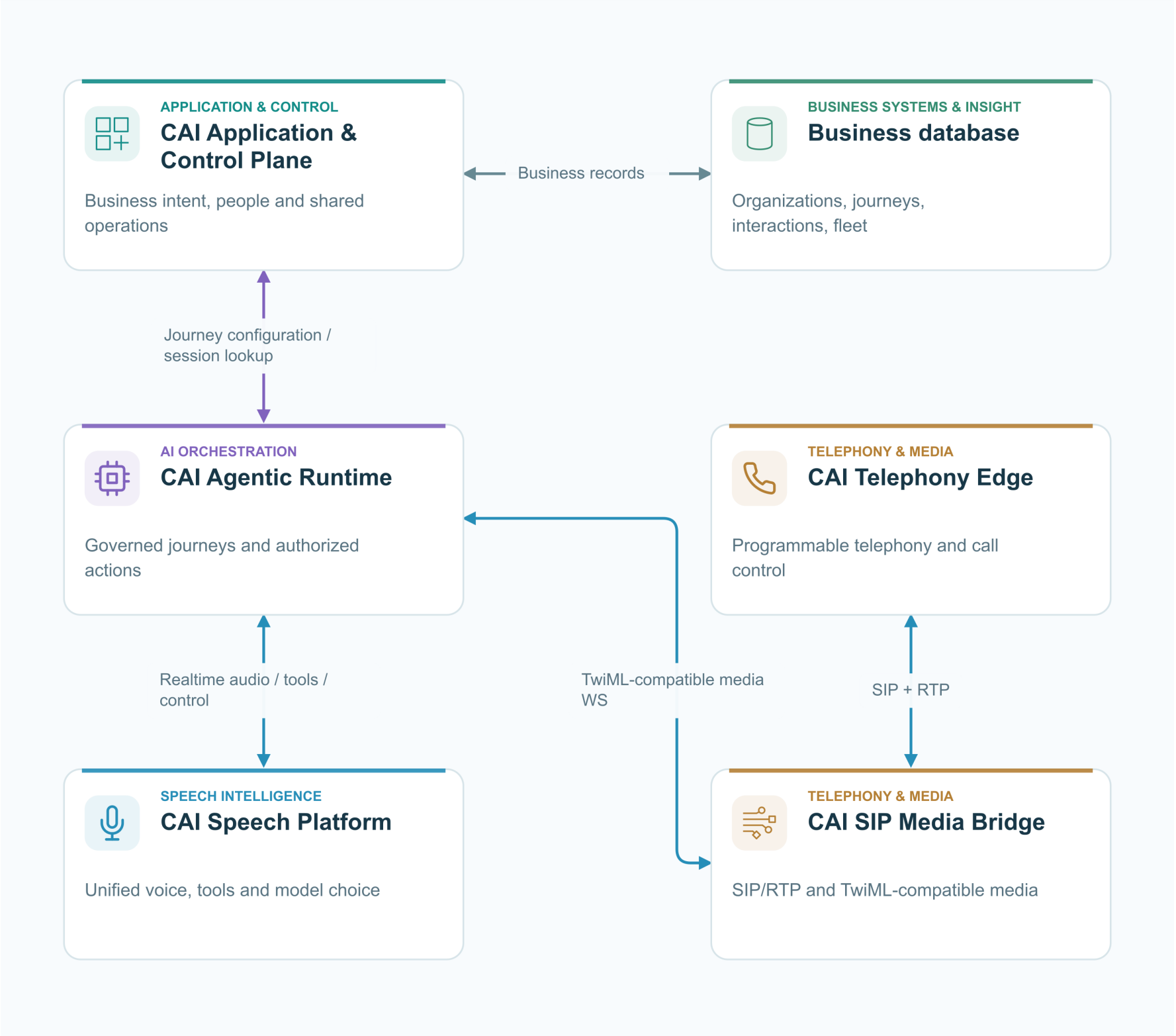
Where does CAI fit?



Business intent and human oversight connect journey design, training and outreach to governed execution. CAI Agentic Runtime coordinates speech and business actions; telephony services connect callers and existing systems. Interaction records return to the application and Fleet Control, where reviewed improvements feed back into journey design. These connected layers form one supervised customer operation.

D02 · Five building blocks

Who owns what?



CAI Application & Control Plane owns business configuration and durable records. CAI Agentic Runtime owns the live journey and authorized actions. CAI Speech Platform provides the common voice interface and engine selection. CAI Telephony Edge controls programmable call legs, while CAI SIP Media Bridge adapts media. WebRTC and direct speech clients can use CAI Speech Platform independently.

Responsibilities & planes

CAI carries conversations through its media services, governs actions through the runtime and records outcomes in shared business records. Speech and DTMF flow between callers and AI; journey configuration and policy determine what the system can do; events and recordings support reporting and improvement.

The application provides the runtime with journey/revision and organization-derived speech configuration. The runtime conducts the live interaction and sends selected effects to the speech and telephony layers. Events and post-call work return durable evidence to the application's records.

The application contains multiple workspaces over shared business domains. The speech service keeps its own configuration and identity store. Distributed deployment does not merge these authorities into one database or require each component to occupy its own machine.

Media, control and operational records

Plane	Examples	Question answered
Media	SIP/RTP, compatible media WS, speech realtime WS, WebRTC	Where does the conversation travel?
Control	Journey lookup, profile resolution, tool authorization, ESL transfer	Who is allowed to decide or change something?
Evidence / data	Interaction events, recording artifacts, feature records and findings	What happened, where is it retained, and how can it be attributed?

Journey configuration selects behavior, transfer commands control call legs and transcripts record the conversation. The runtime maintains authoritative journey state.

Application root

Organizations and users define the access and configuration context. Journey Studio supports authoring and revision workflows. Training maintains learning/simulation records. Outbound Campaigns configures contact attempts. Fleet Control turns operational evidence into issues, proposals and reviewed changes.

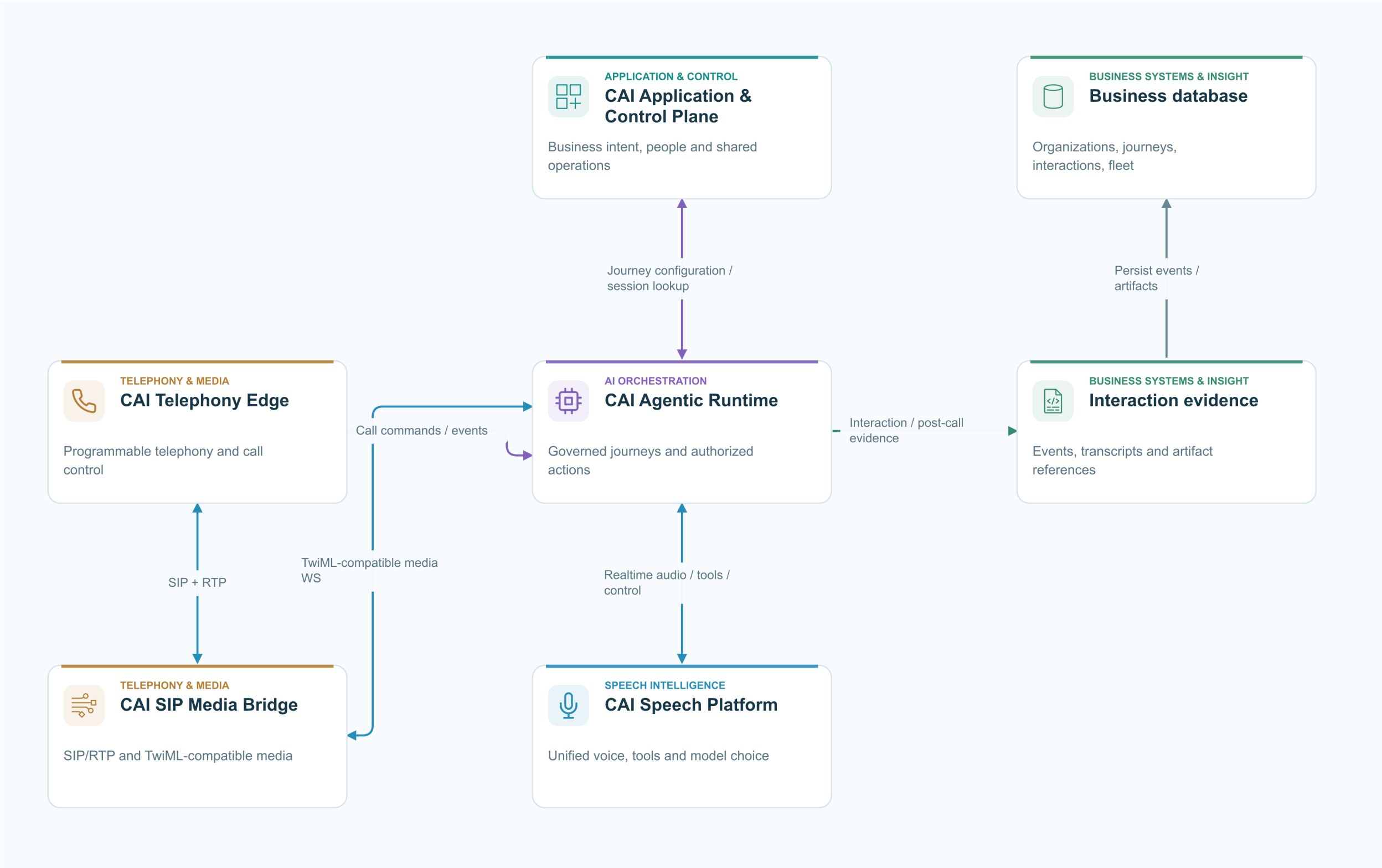
The runtime and speech services are execution partners of this root. Their transient state, worker queues and speech configuration have explicit ownership. The app-to-runtime speech configuration endpoint derives organizational ownership server-side; a client-supplied organization label is not sufficient authority.

Interaction identity

Call, stream, journey revision and interaction identifiers connect the live conversation to its business records. Journey deployment bindings select the behavior for an entry point; cloud placement determines where the services execute.

D03 · Media, control and evidence

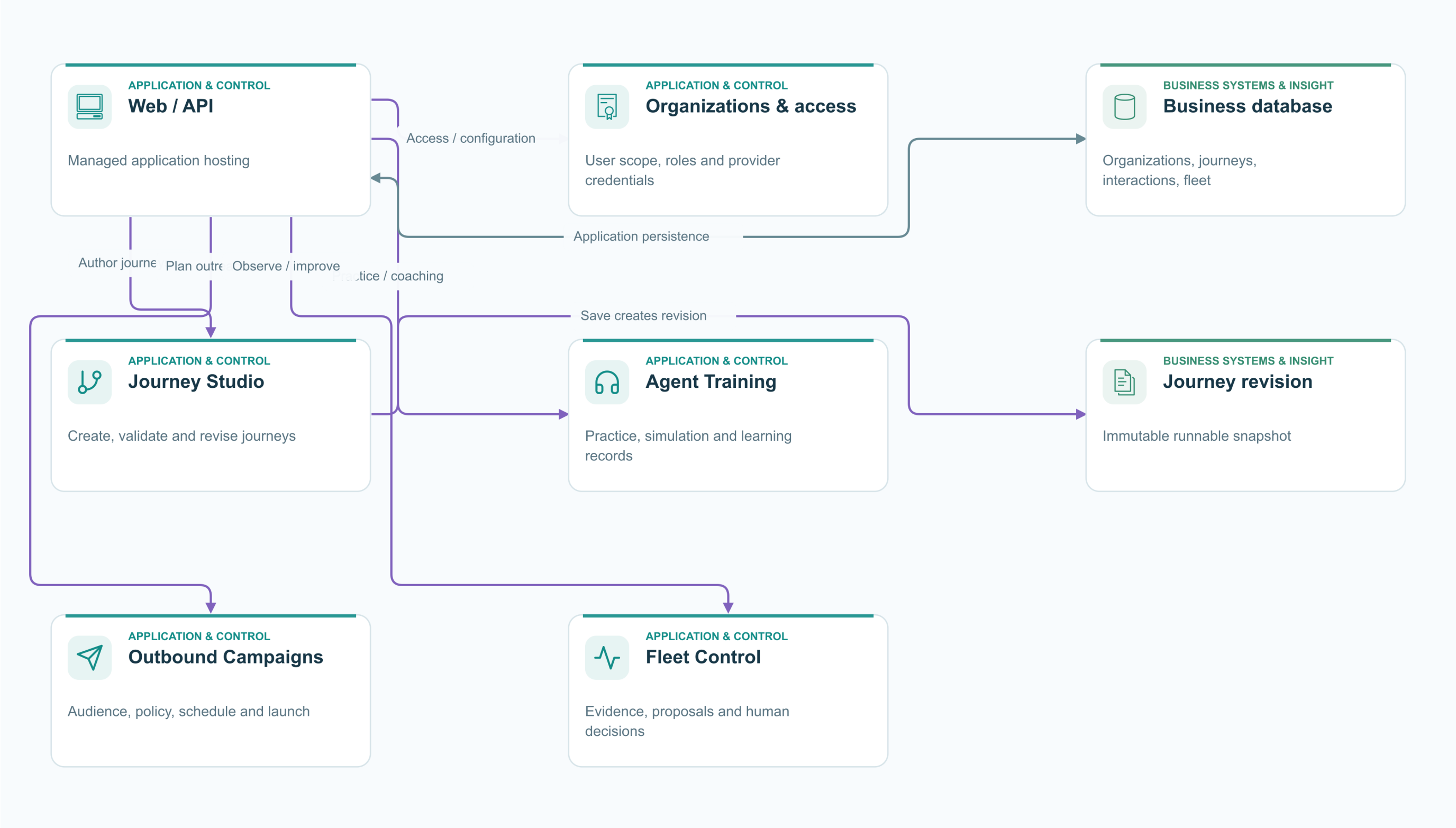
What flows where?



Media carries the conversation between telephony, CAI Agentic Runtime and CAI Speech Platform. Configuration and call-control exchanges govern journey behavior and telephony actions. Interaction events and post-call results return to the business database. The application supplies configuration and receives outcomes while audio flows through the execution and media services.

D04 · Inside the application root

What does the application govern?



CAI Application & Control Plane connects user-facing workspaces to shared business records. Journey Studio produces versioned runnable journeys; training maintains practice and learning records; outbound campaigns coordinate outreach; Fleet Control manages observed issues and reviewed improvements. Organizational access and configuration connect these experiences across the operation.

Governed execution

The runtime makes the journey an application-owned process. A speech model interprets and generates, while controllers own turn-taking, phases, slots, guards, allowed tools and consequential state changes. A model's function call is a request for an action, not a grant of authority.

CAI Agentic Runtime provides turn, journey and policy control with DTMF, structured input and human influence. Controllers enforce the journey and action policy throughout the live conversation.

Preflight prepares the interaction before conversation starts. Closeout records the outcome and triggers post-call processing, connecting live execution to the business record.

A representative lifecycle

1. A compatible call entry supplies identity and media metadata.
 2. The runtime resolves the scenario/revision and runs bounded preflight.
 3. For a CAI speech journey, the app resolves organization-owned endpoint, profile and credential material.
 4. Controllers manage listening, response release, interruption, tools and journey progression.
 5. Human approvals, takeover or telephony transfer use structured control paths.
 6. Closure stops the interaction and triggers evidence/post-call processing.
- Speech configuration failures produce an explicit session failure. Provider substitution follows the configured policy.

State and authority

Mechanism	Responsibility
Turn controller	Release responses, handle speaking/listening and interruptions
Journey controller	Commit phases, slots, guards and progress
Policy engine	Determine whether actions require approval or must stop/escalate
Tool runtime	Execute allowed actions and normalize public/private/frontend results
Human control	Convert approval, override and takeover into explicit events

Human takeover and telephony transfer are distinct. The first changes interaction authority; the second changes call legs. Some scenarios require both.

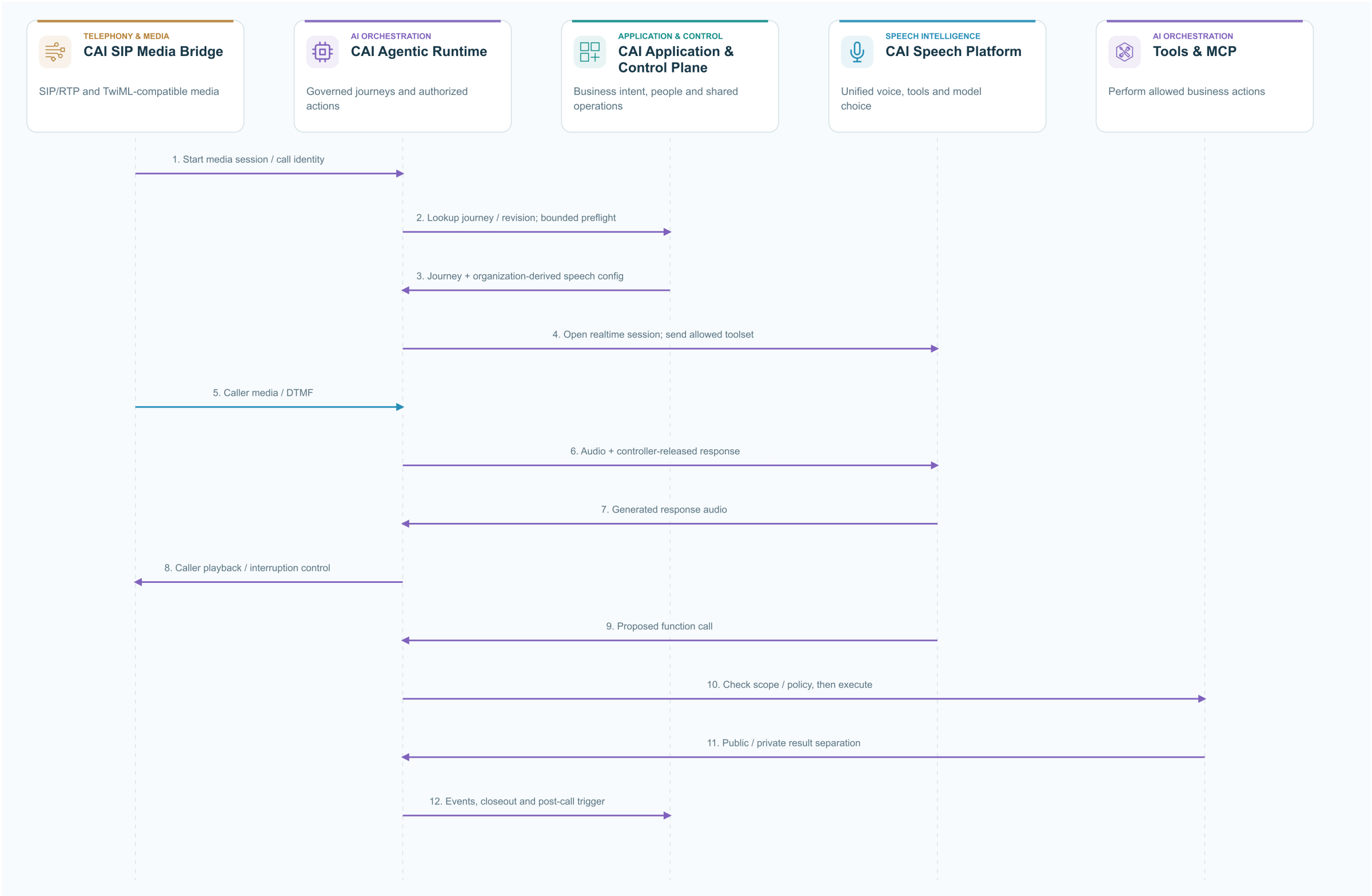
Failure ownership

Provider errors belong to session execution and its chosen fallback policy. Tool errors must produce a controlled outcome without silently advancing an invalid journey. Denied approval must not look like successful execution. An interrupted response must be reconciled with actual playout, particularly for recording and final-state evidence.

Journey validation covers successful completion, approval rejection, preflight failure, transfer and provider or tool errors.

D05 · One governed inbound call

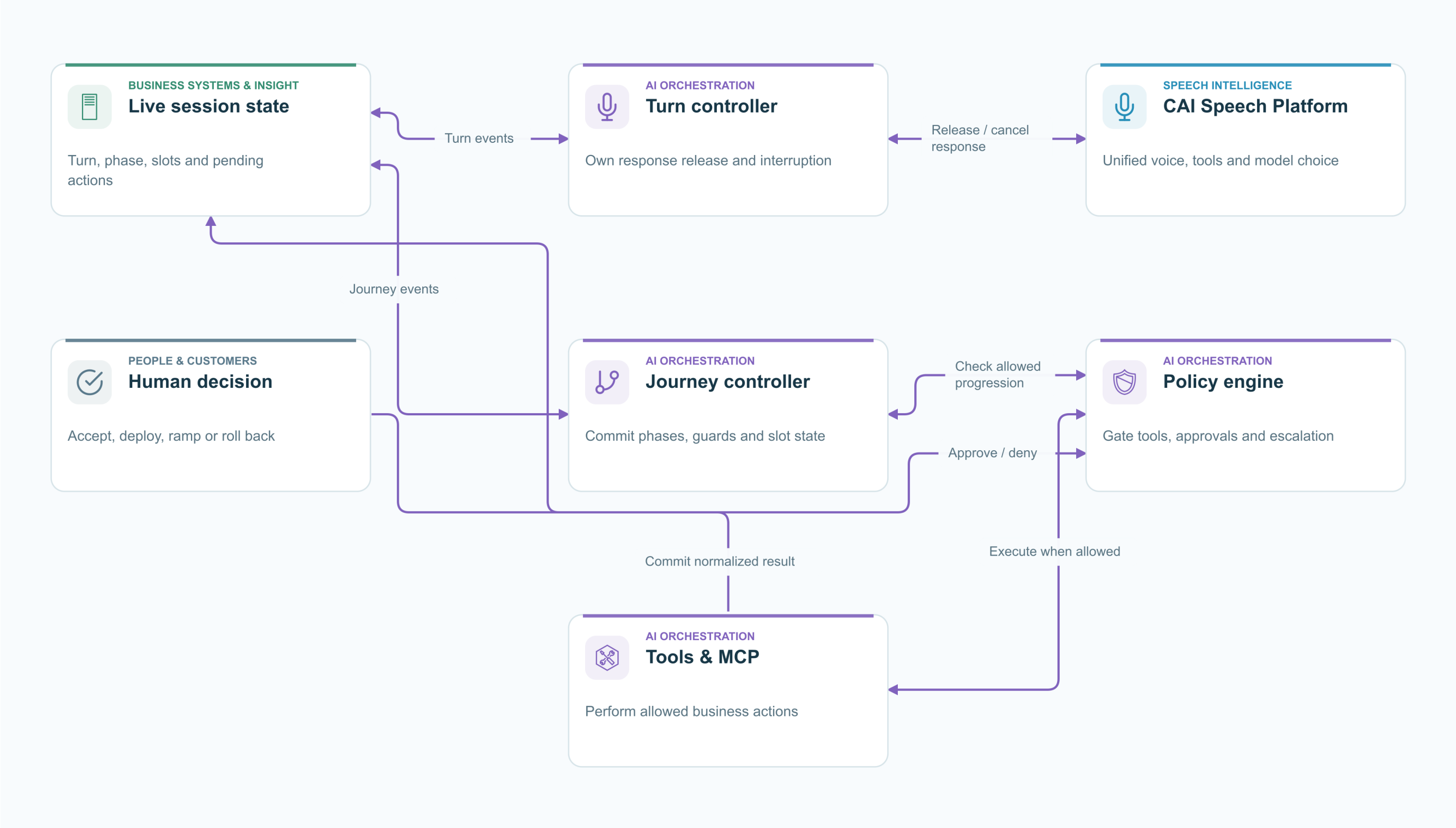
What happens during a call?



The production call path enters CAI Agentic Runtime through the compatible media interface. Startup resolves the journey and speech configuration. During conversation, controllers govern response release and tool execution. Selected tool results return to the model, and closeout connects the interaction to post-call processing and business records.

D06 · The runtime owns authority

Which decisions belong to the application?



Provider and human input enter the runtime state machine. The turn controller governs speaking and interruption; the journey controller governs phases, guards and slots; policy gates consequential actions. Tool execution returns normalized outcomes into state. A model can propose a function call, while runtime policy determines whether it executes.

Speech & WebRTC

CAI Speech Platform provides a unified speech interface with interchangeable engines. A profile selects native speech-to-speech, a composed STT → LLM → TTS pipeline, or a hybrid engine combining speech understanding with independently selected synthesis.

Realtime WebSocket and in-house WebRTC are two interfaces within CAI Speech Platform. Clients can use either directly, or use CAI Agentic Runtime for governed journey execution.

The speech pipeline supports denoising, voice enhancement, voice activity detection, recognition, language reasoning and synthesis. Tools and MCP connect conversations to external capabilities.

Common interface, flexible engines

The realtime WebSocket interface uses `/v1/realtime` and a common event vocabulary based on realtime speech interaction. Profiles separate provider credentials and engine settings from client behavior and journey design.

Capability	Function	Configuration
Native speech-to-speech	Combined speech understanding, reasoning and generation	Provider profile and turn settings
Composed speech	Independent recognition, language model and synthesis	STT, LLM and TTS selection
Hybrid speech	Speech understanding with independently selected output voice	Input engine and output synthesis profile
Enhancement and VAD	Denoise, enhance and admit speech input	Processing options supported by the selected engine path
Tools and MCP	Connect conversations to business capabilities	Tool definitions, external execution and returned results
Browser voice	In-house WebRTC media and signaling	Authorized session configuration and optional TURN relay

Audio and turn control

Input processing follows the selected engine profile. Enhancement and VAD prepare admitted audio where enabled. Native and composed engines retain their respective turn behavior; cancellation, context handling and reconnect behavior follow the engine contract.

Output combines audio delivery with playback tracking. Interruption cancels or truncates the relevant response so conversation state reflects what the caller actually heard.

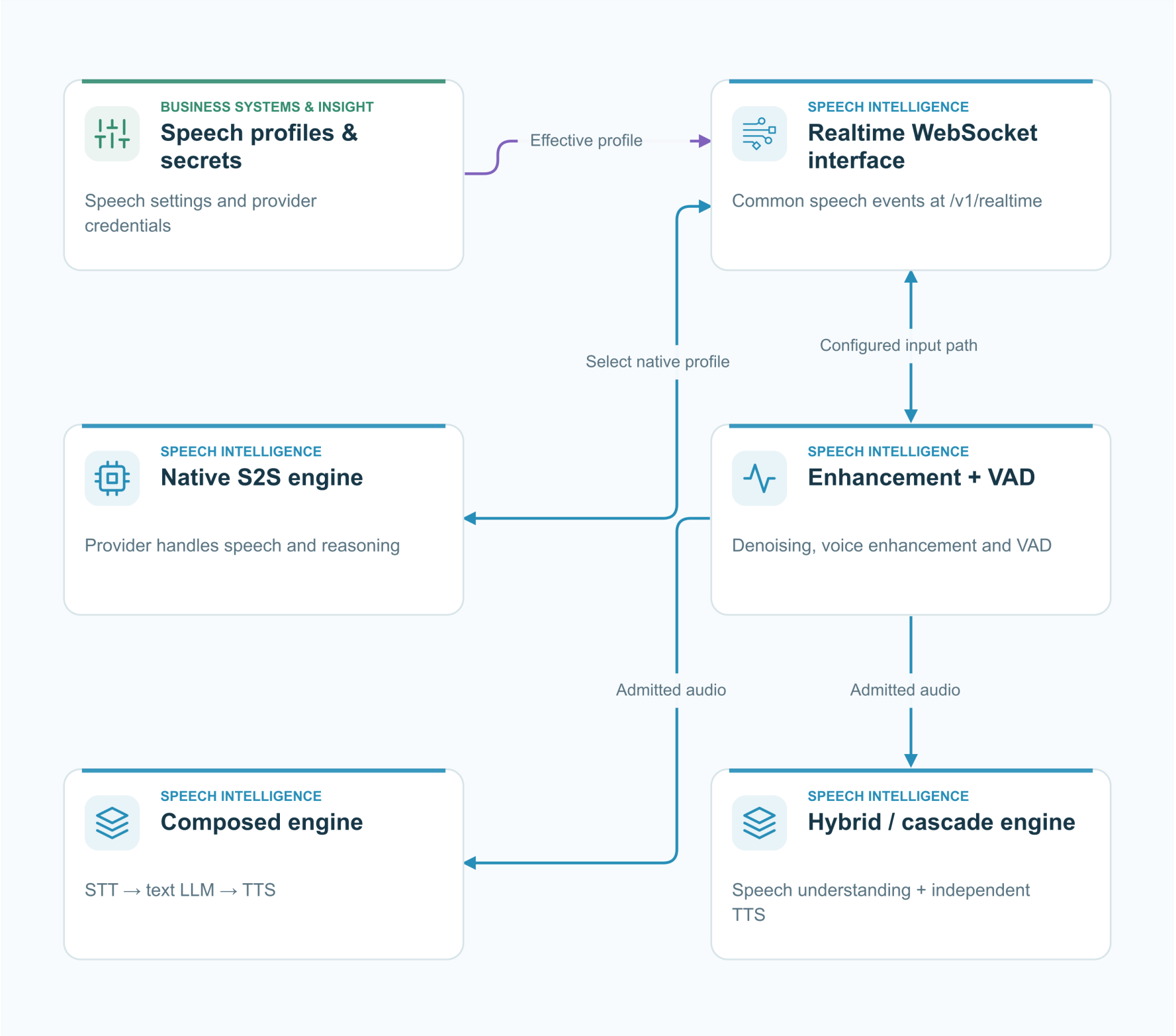
Browser sessions

The application authorizes a training session and resolves organization-specific configuration. The WebRTC interface handles browser media and adapts speech events to the realtime interface. TURN provides network traversal where a direct connection is unavailable.

CAI Speech Platform maintains speech profiles, service identities and encrypted provider credentials in its configuration domain. Independent clients manage their own business workflow; integration with CAI Agentic Runtime adds journey policy and controlled actions.

D07 · One interface, several speech engines

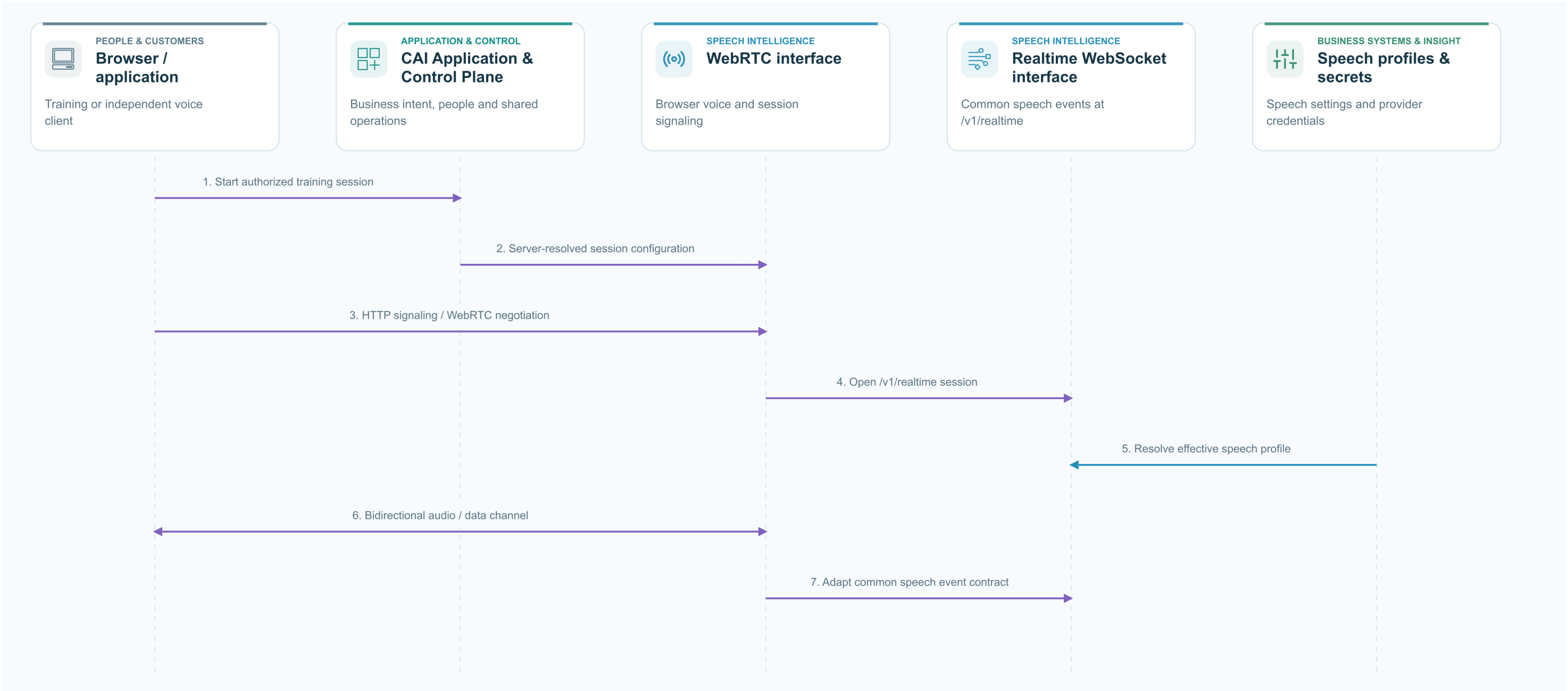
How does model portability work?



A speech profile selects the engine behind the common realtime interface. Native speech-to-speech handles speech and reasoning together. Composed engines separate recognition, language reasoning and synthesis. Hybrid engines combine speech understanding with independently selected output voices. Enhancement and VAD follow the selected processing path, while clients retain the common speech event contract.

D11 · Browser voice and training

Where does WebRTC live?



The application authorizes the training session and resolves organization configuration. Browser signaling and media reach the WebRTC interface within CAI Speech Platform, which adapts events to the common realtime interface. Speech profiles select the engine and settings. An optional TURN relay supports network traversal for browser media.

Tools & sensitive data

CAI Agentic Runtime separates trusted execution state from model-visible results. A tool can receive private input, perform a business operation and return a selected outcome, such as verification accepted or failed. The model can continue the journey without receiving the underlying private fields. Action policy and data visibility work together. Policy governs whether an operation may run; visibility governs which inputs and results reach the model, a human workspace or an operational record.

Information boundaries

Audience	Information	Control
Trusted runtime and tool	Inputs required for the operation	Scoped execution context and authorization
Speech model	Selected conversation results	Explicit public result projection
Human workspace	Customer and operational context appropriate to the role	Organization and user access
Audit and operational records	Necessary action and outcome traces	Configured access, retention and redaction policies

Private collection paths are selected before sensitive data enters the conversation. Audio sent to a speech provider is already available to that provider; later result filtering does not remove earlier audio. DTMF, trusted forms and tool-side verification offer distinct collection paths with their own recording and retention configuration.

Authentication and authority

The application resolves organizational membership and server-side provider credentials. Runtime observer access uses session authorization, and speech clients use authorized profiles and session grants. Service identities and human approvals have separate roles in the execution record.

Tools and MCP

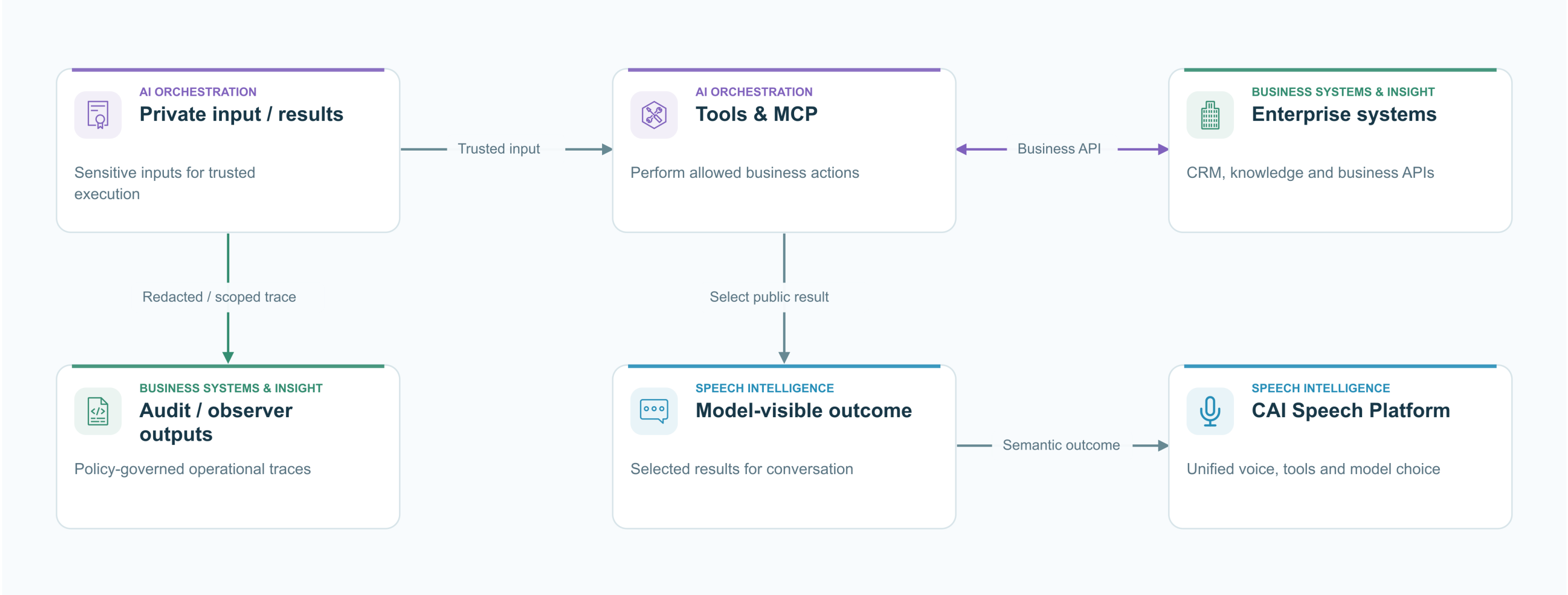
CAI supports function tools and MCP integrations. A conversation requests a tool through the common realtime interface. CAI Agentic Runtime evaluates scope and policy, invokes the authorized business or MCP tool, and returns the selected result to the conversation. Independent speech clients can connect their own tool executor. In a governed journey, CAI Agentic Runtime retains authority over actions and journey progression.

Verification example

A trusted verification tool receives a private identifier and checks it against the business system. The model receives the verification outcome and the permitted next step. Authorized operational records retain the action and result according to the configured access and retention policy.

D08 · Private inputs, public outcomes

Who can see the result?



Trusted tool execution retains private inputs and returns selected outcomes to the conversation. The model receives the information needed to continue the journey. Authorized human workspaces and operational records have their own access and retention policies. Collection channels determine which services receive sensitive information before processing begins.

Connection choices

CAI can attach at several layers. The production chain connects existing PBX telephony through the SIP/media bridge and governed runtime to the CAI speech platform. Other clients may enter through a narrower SIP path, a provider-supported media path, a compatible runtime WebSocket, a direct speech WebSocket or in-house WebRTC.

Choose the attachment point according to the responsibilities the existing ecosystem should retain. Keeping a customer ACD and desktop is compatible with using CAI journeys, training and fleet improvement. Owning the entire operation adds queue/presence/desktop responsibilities; it is not just another audio transport.

CAI SIP Media Bridge provides TwiML-compatible media connectivity directly. Twilio is an optional provider path.

Connection options

Pattern	Keep / supply upstream	CAI entry
P1: PBX via FreeSWITCH	Carrier/PBX, numbers and human destinations	CAI Telephony Edge, CAI SIP Media Bridge and CAI Agentic Runtime
P2: Direct native SIP	SIP peer and any omitted call-control responsibilities	CAI SIP Media Bridge; PCMU/RTP and telephone-event DTMF
P3: Twilio-supported	Provider voice service	Runtime TwiML/media routes; also current outbound dialer path
P4: Runtime media WS	Compatible media client/adaptor	Journey runtime; requires event envelope, metadata and encoding, not simply any WebSocket
P5: Direct speech WS	Client's business process and tool execution	Speech /v1/realtime ; independent of the journey runtime
P6: Browser WebRTC	Browser and authorized session broker	In-house WebRTC component inside CAI Speech Platform

Telephony protocols and call control

SIP handles signaling; RTP carries media; ESL controls FreeSWITCH. TwiML HTTP instructions and Twilio-style media WebSocket events are different contracts. The speech gateway uses another realtime event vocabulary behind the runtime adaptor.

The bridge correlates media identity with the FreeSWITCH UUID and call/session identifiers. SIP/RTP carries audio and telephone-event DTMF; the media WebSocket carries the compatible start, media and control events. Transfer commands use the correlated telephony identity.

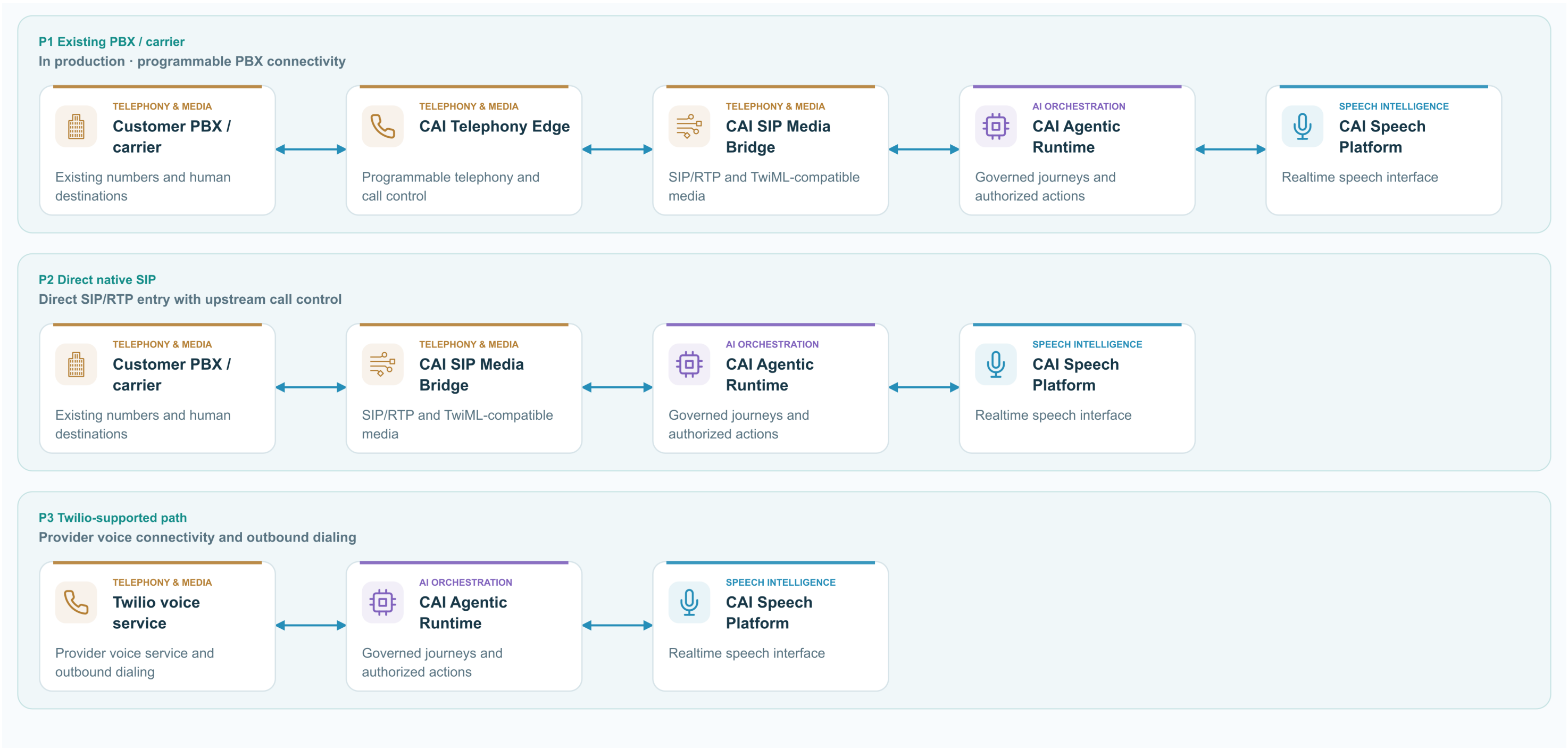
Transport protection is configured per connection: SIP/TLS and SRTP depend on the telephony endpoint, while secure WebSocket protects the compatible media and speech interfaces. Media encoding, playback completion, interruption and recovery are part of each connection contract.

Decide the operational owner

For each pattern assign carrier/trunk, numbering, queue/routing, agent presence, call control, model configuration, business record, recording retention, incident response and change approval. Removing FreeSWITCH from a route requires another owner for its omitted behaviors. Define the fallback destination, hangup/transfer recovery and caller-ID expectations before calling two paths interchangeable.

D09 / 1 · Connection choices: telephony

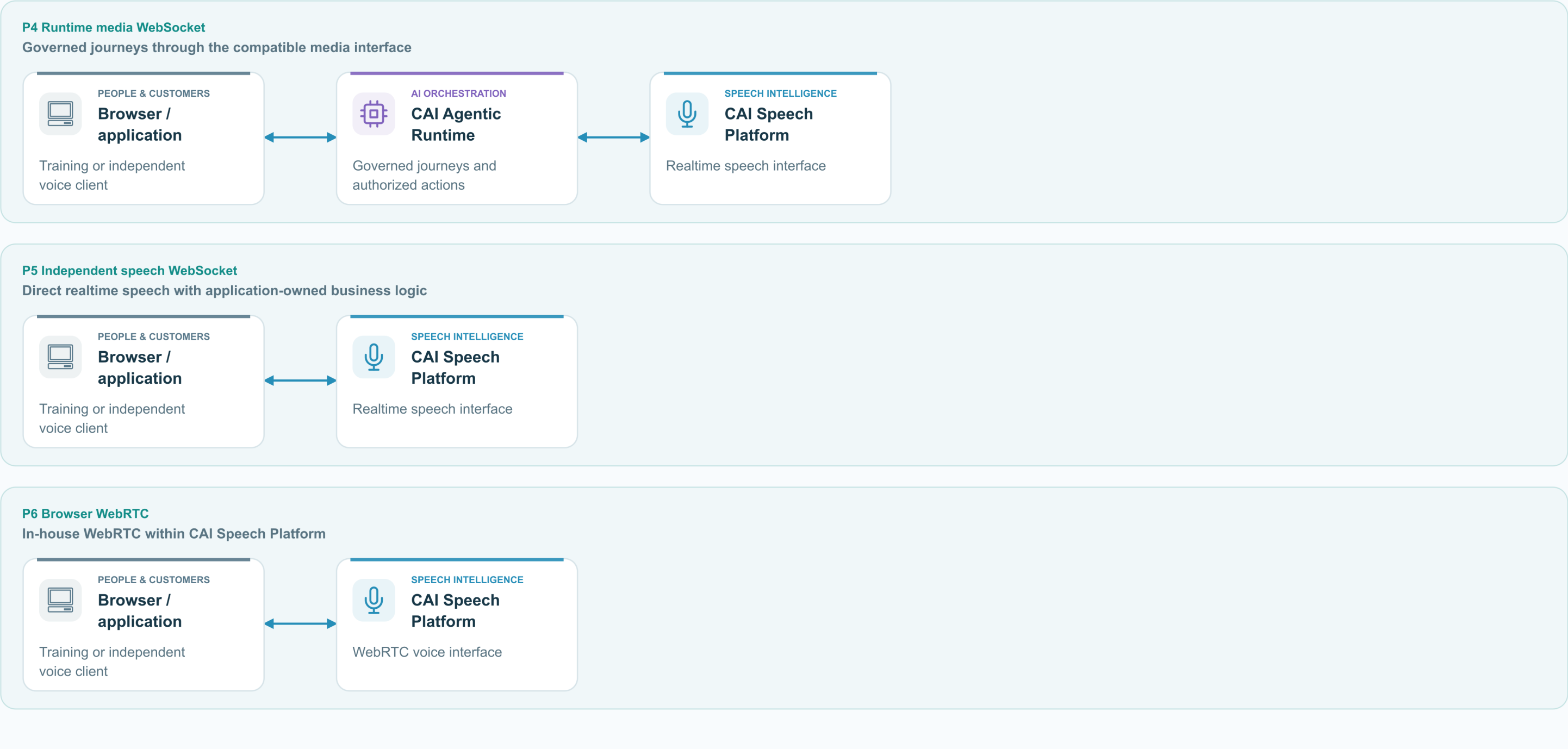
Where can you attach?



CAI offers six attachment choices: a PBX through CAI Telephony Edge, direct SIP through CAI SIP Media Bridge, Twilio, a runtime media WebSocket, a direct speech WebSocket, or browser WebRTC. Organizations retain the systems they need and choose where CAI takes responsibility. Both speech interfaces belong to CAI Speech Platform.

D09 / 2 · Connection choices: WebSocket and WebRTC

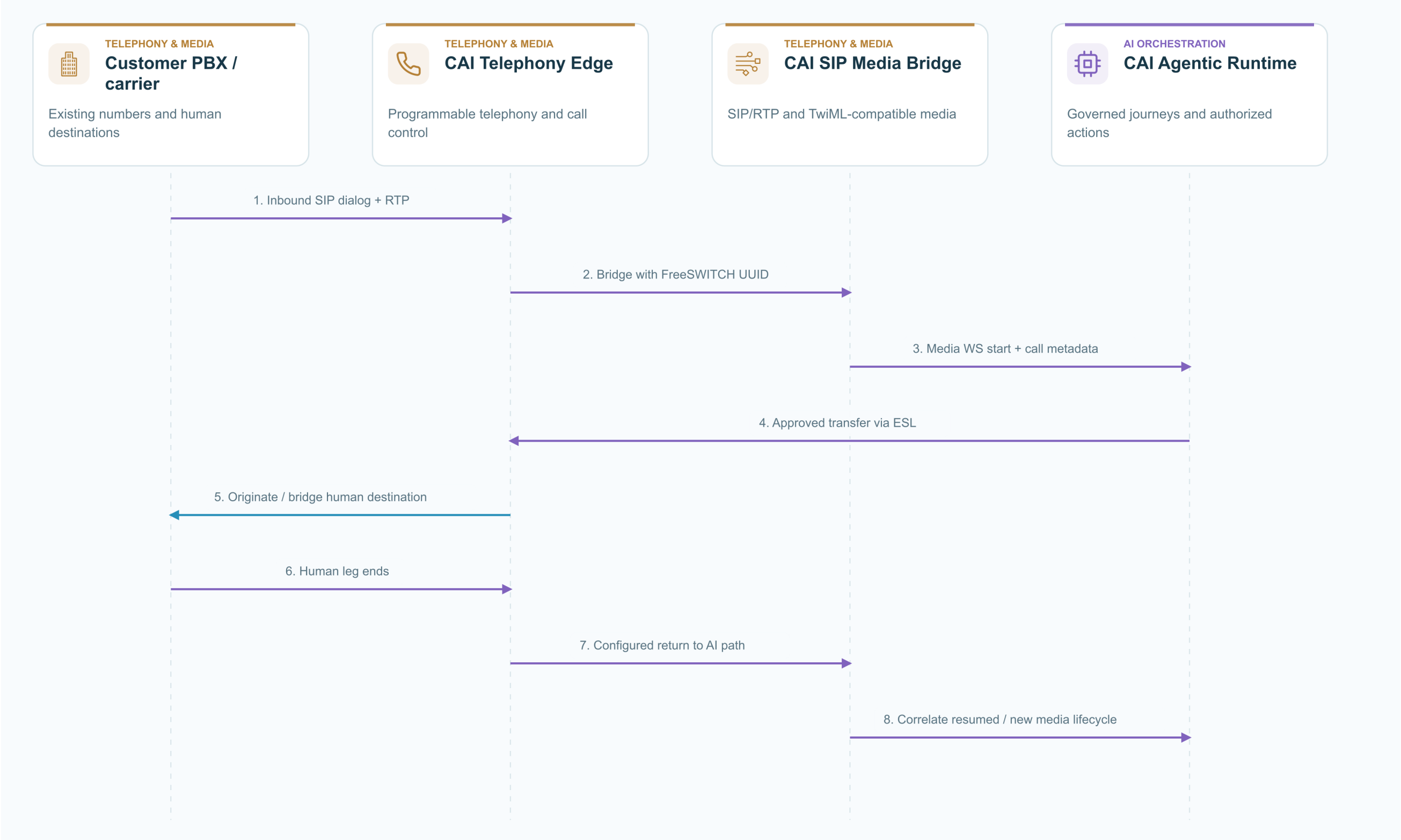
Where can you attach?



CAI offers six attachment choices: a PBX through CAI Telephony Edge, direct SIP through CAI SIP Media Bridge, Twilio, a runtime media WebSocket, a direct speech WebSocket, or browser WebRTC. Organizations retain the systems they need and choose where CAI takes responsibility. Both speech interfaces belong to CAI Speech Platform.

D10 · Telephony transfer and return

Who owns the call legs?



Media adaptation and telephony call control are different responsibilities. The bridge carries compatible media messages to the runtime. When the journey authorizes a transfer, the runtime can control FreeSWITCH through ESL using the correlated FreeSWITCH UUID. The carrier/PBX leg reaches the human destination. Return-to-AI behavior is a configured dialplan behavior, not a generic guarantee of every direct SIP integration.

Outbound execution

An outbound campaign starts as app-owned business intent: audience, policy, schedule, dialing settings and journey choice. The runtime receives a frozen launch snapshot and executes eligible attempts through the outbound worker and queueing layer.

The outbound dialer integrates with Twilio. Answered calls use separate outbound runtime entry points and then governed conversation behavior. Inbound SIP bridge flexibility does not automatically give the outbound subsystem equivalent carrier-native dialing.

Execution state exists in more than one place: app campaign/attempt records, worker queues and runtime snapshots. Callbacks return attempt outcomes to the application. Reconciliation connects provider status, worker progress and campaign records when delivery is delayed or retried.

Attempt lifecycle

A campaign planner selects eligible contacts, considers timing/policy and reserves attempts. BullMQ coordinates queued or delayed jobs. The dialer initiates calls. Runtime handling records status and conversation disposition. Attempt/campaign events return to the app, and retry policy decides whether another attempt is allowed.

The app brokers management operations such as start, pause, resume and cancel. The application authorizes these operations before they reach internal runtime management routes.

State and recovery

State	Owner	Recovery question
Campaign/list/policy	App + business database	Which frozen configuration was launched?
Queued/delayed jobs	Redis/BullMQ	What work remains after worker restart?
Attempt snapshot	Runtime state store	Was an attempt reserved, dialing or terminal?
Provider call	Telephony provider	Did the call really start/answer/end?
Reported outcome	App callback ingestion	Did the callback arrive, and does it reconcile?

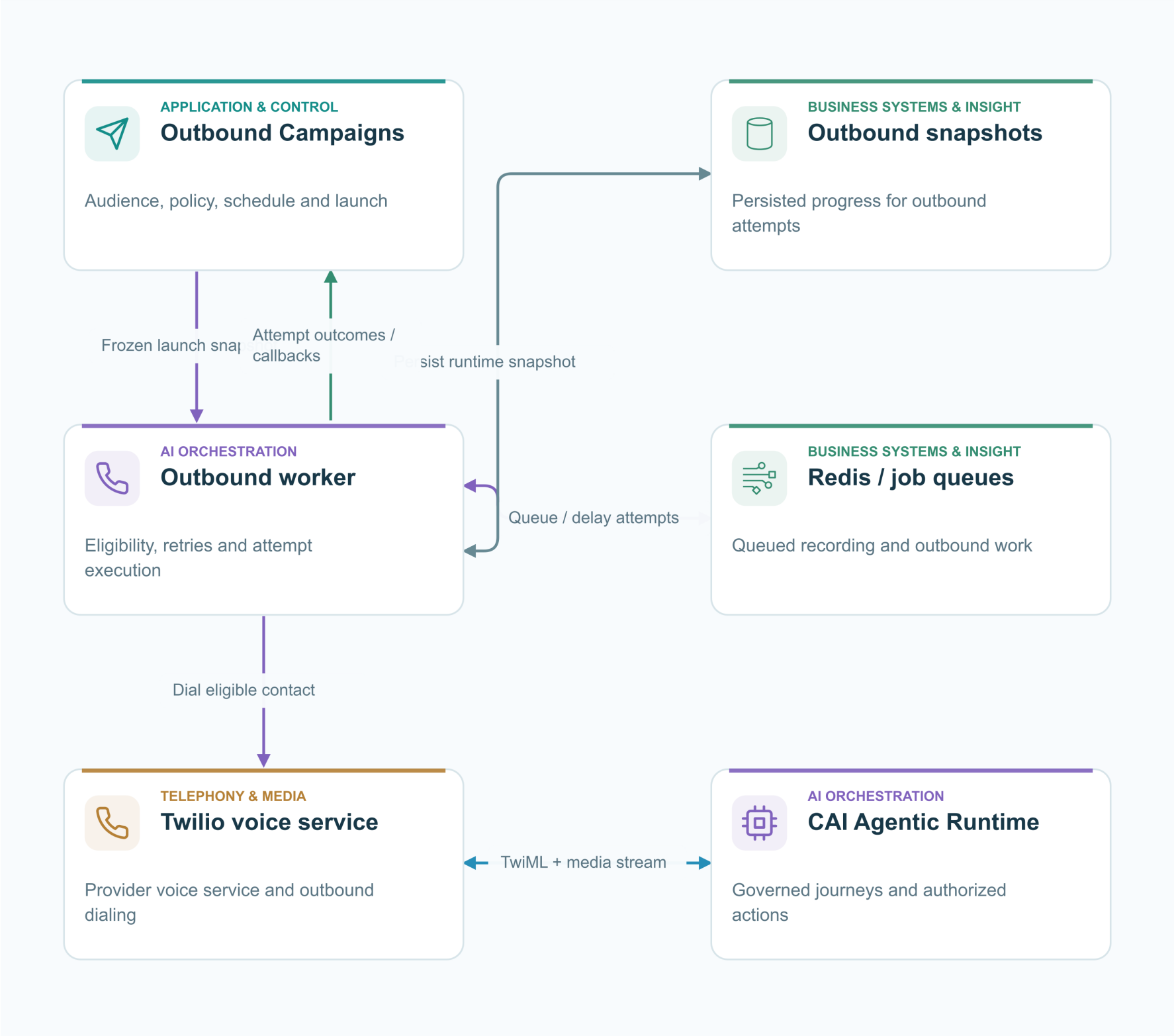
Operational controls include pause and cancel, duplicate prevention, retry limits, restart recovery and terminal-state reconciliation.

Distinct from fleet campaigns

An outbound campaign conducts outreach. A fleet improvement campaign scopes issues, interventions and experiments. Their objects can relate to the same journeys and interaction evidence, but their authority and lifecycle are different.

D12 · Outbound campaign execution

How does a plan become calls?



The application launches a frozen campaign snapshot. Runtime planning selects eligible attempts, with Redis and BullMQ coordinating queued and delayed work. Persisted snapshots support restart recovery. Twilio initiates outbound calls and connects answered conversations to CAI Agentic Runtime. Attempt callbacks and reconciliation connect worker progress to campaign reporting.

Fleet improvement

Fleet Control turns interaction evidence into a reviewed improvement cycle. Each deployment-resolved call can carry a revision assignment. Post-call processing reduces interaction events into a feature record. Watches detect patterns, issues hold the evidence, and a person requests and reviews a proposed intervention.

The primary example is the **Spitch FAQ** program. Repeated questions the journey cannot answer create a knowledge gap. A candidate answer/revision is tested against the baseline, with human decisions controlling rollout.

Walk the FAQ example

1. Bind a number to the journey's production ref or a controlled experiment.
2. Synthetic/adversarial callers ask a deliberately unsupported question.
3. Feature extraction records unanswered topics and appropriate outcome/probe signals.
4. A watch raises an issue with examples and affected calls.
5. A person requests a proposal, supplying or checking the correct answer.
6. The system produces a detached candidate, a reviewable diff and a draft experiment.
7. A person starts the experiment and later chooses ramp, finalize or rollback.

Attribution and authority

Immutable journey revisions separate what ran from what is currently being edited. Refs move under controlled actions; experiment arms pin revisions. Assignment stamps connect observed outcomes to the decision that routed the call.

The app owns watches, findings/issues, candidate records and decisions. The runtime fleet worker is a scheduler that asks the app to evaluate due watches. It does not become the business system of record.

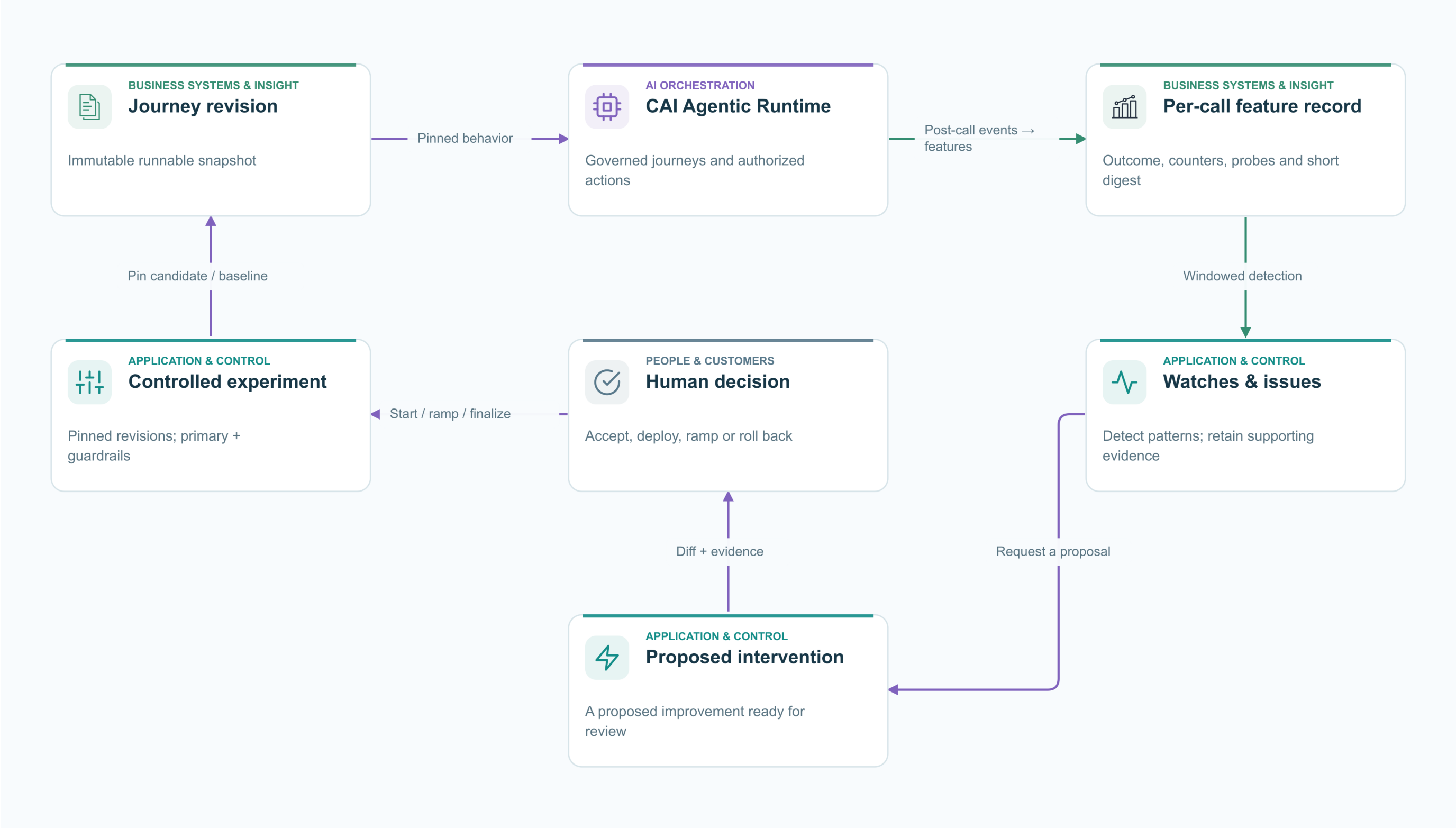
The program's stance is detect and propose. Human actions control caller-impacting changes. Claimed external actors, executing program versions and authenticated approving users must not be collapsed into one audit identity.

Verification

The FAQ experiment uses a primary hypothesis outcome and error/handoff guardrails. 'Not asked' is not the same as 'unanswered.' Insufficient evidence remains a valid result. The program can record no action, dismissal or rollback rather than requiring every detected issue to produce a rollout.

D13 · The Spitch FAQ improvement loop

How does evidence improve a journey?



The Spitch FAQ example demonstrates a complete fleet improvement cycle. Repeated unanswered questions produce feature records and an issue. A person reviews a proposed answer and starts an experiment. The comparison uses a primary outcome with error and handoff guardrails. Human decisions control ramp-up, finalization and rollback.

Data ownership

The application is CAI's principal business system of record. It stores organizations, training/simulations, journey definitions and revisions, interaction evidence, outbound business data and fleet improvement objects.

Other stateful domains remain distinct. The runtime owns live session state, Redis/BullMQ jobs, outbound snapshots and recording-processing files. The speech platform owns independent users, provider configurations, API keys and runtime profiles.

Recording references link the business record to media at its configured storage destination. Each service retains clear responsibility for identity, persistence, retention and recovery.

Ownership table

Data domain	Authoritative owner	Important consumers / boundary
User and organization context	Application business database	Workspaces, service configuration and authorized brokers
Journey and immutable revision	Application	Runtime receives runnable behavior and attribution
Live turn/phase/slot state	Runtime controller	Model and UI receive selected projections
Outbound intent	Application	Runtime receives a frozen launch snapshot
Runtime outbound progress	Worker queues + runtime snapshots	Reconcile with provider and app callback records
Speech profile/provider secret	Speech configuration domain; app stores organization connection material	Server-resolved configuration/grant boundary
Interaction event/feature	Application evidence domain	Fleet watches, reporting and authorized observers
Recording bytes	Runtime processing + configured destination	Application persists artifact references
Fleet intervention/decision	Application	Attribution to program, candidate and human principal

Identity and consistency

Stable interaction, call, stream, FreeSWITCH and revision identifiers link records across services. The planned work-item identity connects queueing, human acceptance, AI assistance and follow-up.

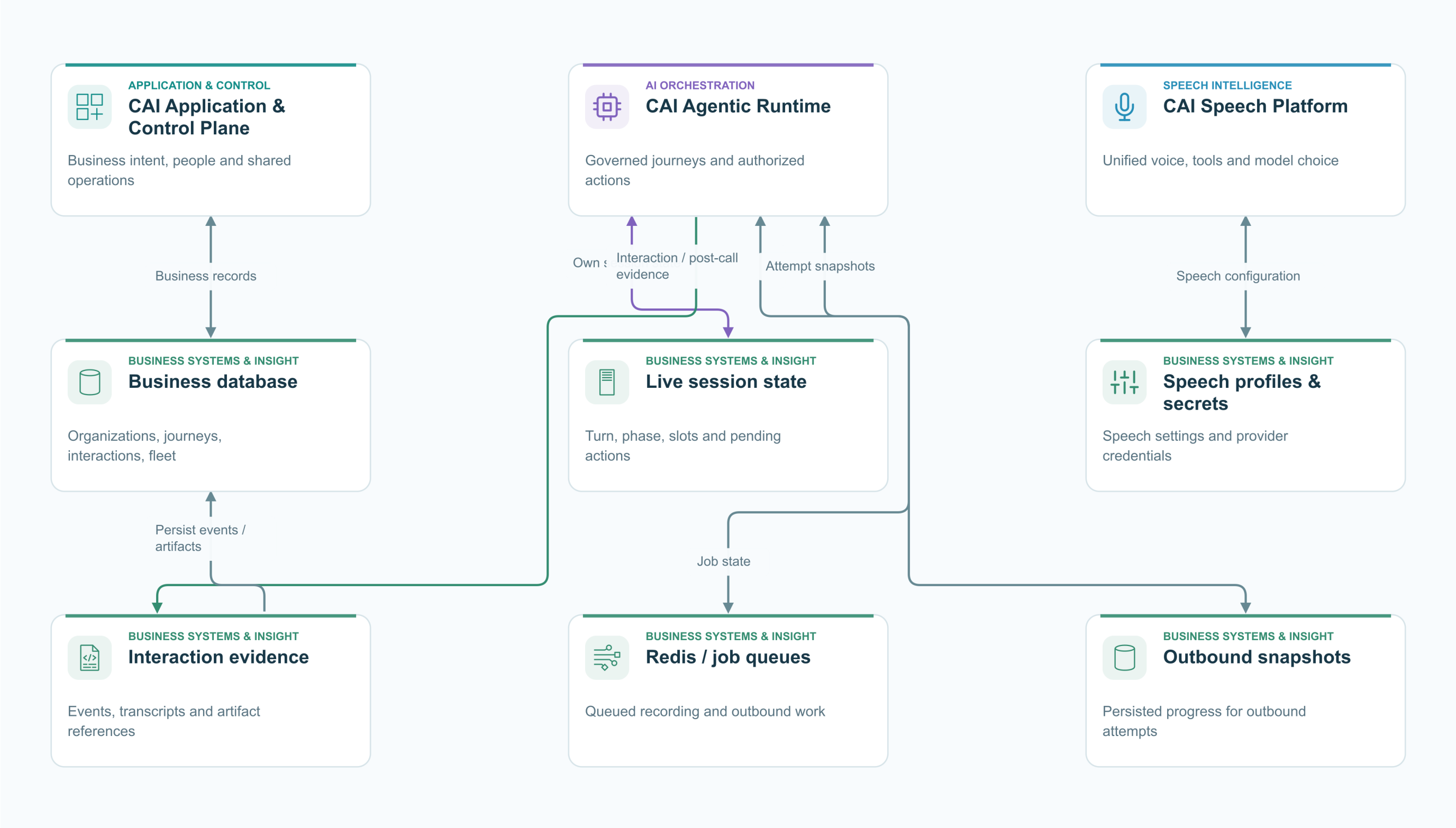
Callbacks and worker retries make idempotency and reconciliation important. Feature extraction and watch evaluation operate on different time scales. Operational projections can lag a live call, and that lag needs to be observable.

Retention and access

Raw inputs, conversation summaries, audit records and recordings have distinct access and retention requirements. Organizational authorization and service-level policies determine who can read each record and how long it is retained.

D14 · State and data ownership

Which record is authoritative?



The business database stores organizations, journeys and interaction records. CAI Agentic Runtime owns live session decisions and worker progress, with Redis supporting queued work and snapshots preserving outbound state. CAI Speech Platform maintains speech profiles and provider credentials. Recording references connect the business record to media at its storage destination.

Deployment & resilience

CAI uses a distributed cloud architecture with managed application and database services and independently deployable execution, telephony and speech services. The application tier can run on Vercel or comparable managed hosting; backend services run in cloud environments such as EC2.

The high-availability strategy uses container-based deployment options: Docker packaging, redundant service instances, health-based routing and controlled rolling updates. Each service can be sized for its workload, separating long-lived conversations from web requests and background jobs.

Components can be colocated for a compact deployment or distributed for independent capacity and fault isolation. WebRTC remains part of CAI Speech Platform in either arrangement.

Deployment tiers

Tier	Services	Operational focus
Managed application and data	Web/API and managed business database	Access, configuration availability, backups and persistence
Runtime execution	Conversation processes, recording, outbound and fleet workers	Session capacity, job recovery and controlled shutdown
Telephony	CAI Telephony Edge and CAI SIP Media Bridge	SIP/RTP reachability, call control and media continuity
Speech	Realtime WebSocket, WebRTC, profiles and engine adapters	Provider availability, audio processing and speech capacity
Adversarial testing	CAI Adversarial Caller	Repeatable scenarios and separation of test traffic

High availability and controlled updates

Redundant instances and health-based routing form the service-availability strategy. Capacity planning considers concurrent calls, audio processing, provider limits and background queue demand separately. The rolling-update design includes session-aware draining: new calls move to healthy capacity while active conversations finish on their owning instance.

Long-lived media sessions require different recovery behavior from stateless web requests. The resilience design includes explicit reconnect or call fallback behavior, recoverable worker jobs and reconciliation of outbound attempt state. Active-call recovery is a session-level concern alongside instance redundancy.

Operational signals

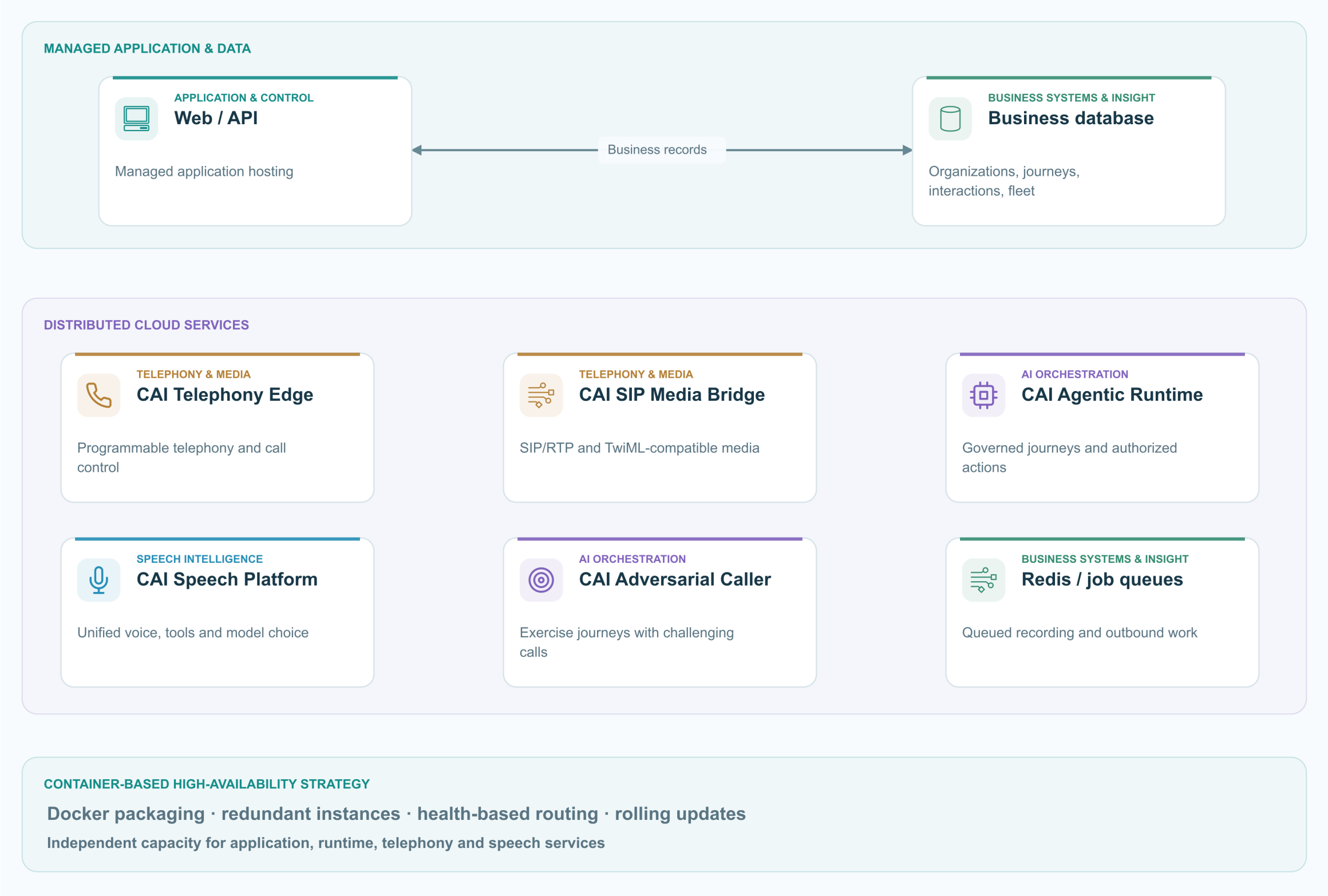
Setup success, call completion, transfer success, speech-turn latency, interruption response, audio continuity and provider errors describe the customer experience. Queue age, event lag, worker recovery and service capacity describe the health of the supporting operation.

Service recovery

Configuration availability governs new session setup. Runtime recovery addresses sessions and actions owned by the affected process. Speech recovery follows the selected provider and journey policy. Telephony recovery follows call-leg and transfer behavior. Background workers recover queued jobs and reconcile persisted progress with business records.

D15 · Cloud deployment & resilience

What runs where?



Managed application and database services connect to distributed runtime, telephony, speech and testing services. Deployments can colocate components or size them independently. The container-based high-availability strategy uses Docker packaging, redundant instances, health-based routing and controlled updates. Session-aware draining and recovery address the long-lived conversations handled by the backend services.

People & operating loop

CAI's operating loop connects business intent to controlled execution and improvement. People design the journey, exercise it, approve rollout, assist where needed and evaluate the evidence. AI conducts supported interactions and helps propose what should improve next.

The same loop connects agent training, journey authoring, adversarial calling, outbound execution and fleet supervision. It is collaborative by construction: observation can be automated, while consequential changes follow explicit authority.

The central management question is not simply how many conversations are automated. It is whether the operation can explain what ran, who controlled it, where intervention is needed and how a proposed change is verified.

Responsibilities by stage

Stage	Human role	System role
Observe	Choose objectives, scope and acceptable evidence	Capture interactions, derive features and detect patterns
Design	Set intent, check behavior and business correctness	Assist authoring and validation
Test	Define cases and judge outcomes	Exercise journeys through examples, harnesses and adversarial calls
Deploy	Authorize revisions, bindings and exposure	Execute controlled routing and record attribution
Assist / coach	Resolve exceptions and develop human performance	Supply context, simulation and selected assistance
Improve	Accept, reject, experiment or roll back	Propose bounded interventions and compare outcomes

Three distinct approvals

In-call approval governs a particular action or transition. Journey rollout approval changes behavior for subsequent interactions. Infrastructure change approval changes the operating estate. A shared UI can expose all three, but the underlying contract and authority must remain distinct.

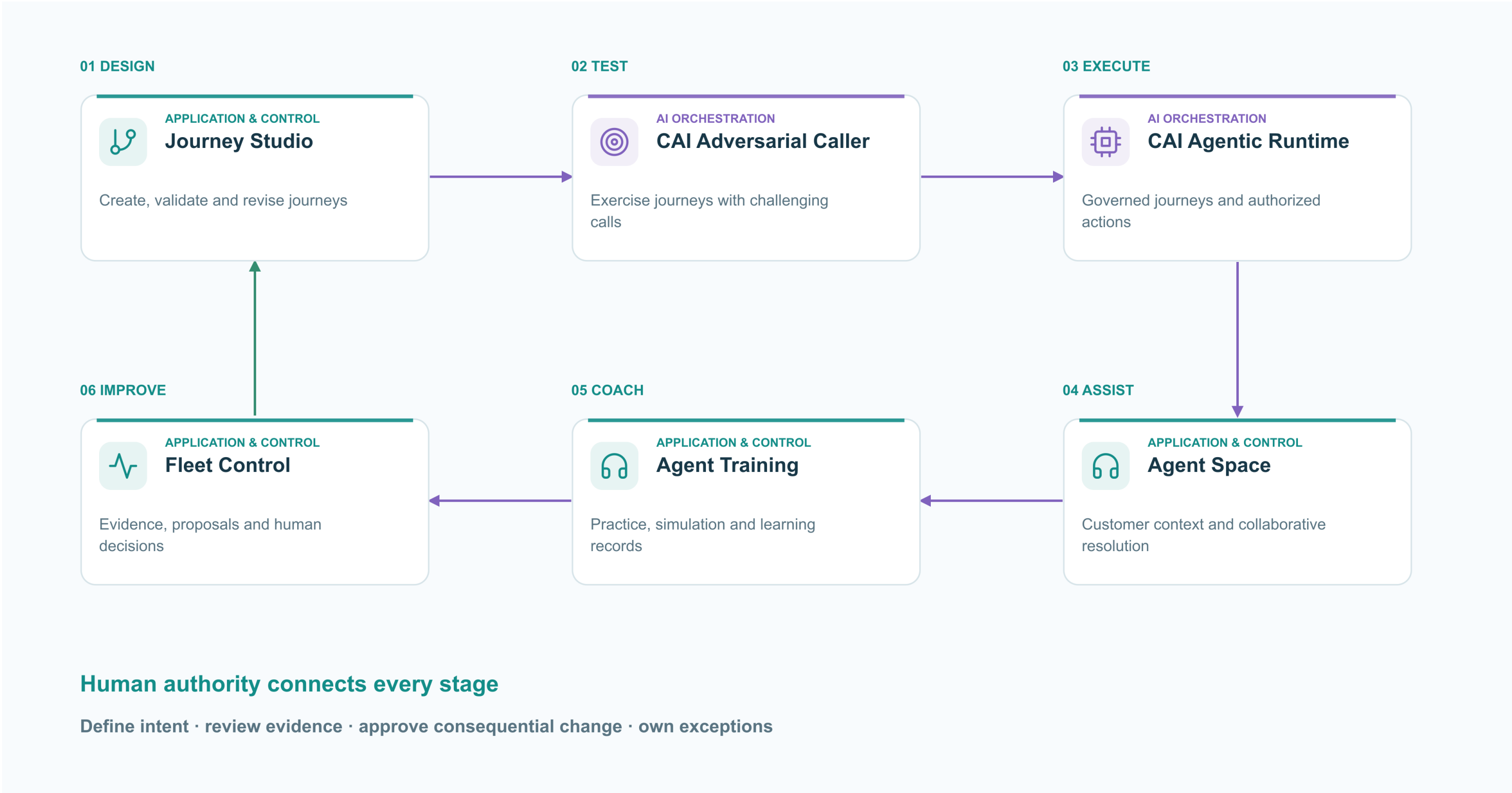
Operational handoffs

Integration agreements should name who owns the carrier/trunk, human queue, desktop, recording destination, business record, incident response and change approval. Existing contact centers can retain most of these while CAI adds governed AI execution and learning.

Full CAI operation adds durable work distribution and human-session ownership. The roadmap connects intake, assignment, resolution and follow-up into one accountable work lifecycle.

D16 · The collaborative operating loop

What do people and AI do?



People define business intent and authorize consequential changes. Journey design and adversarial testing prepare behavior; CAI Agentic Runtime conducts the interaction; assistance and training support human performance. Fleet observations identify recurring issues and proposed improvements. Human review closes the cycle, connecting operational learning to controlled changes in customer experience.

Product workspaces

The Command Center brings the customer operation into one connected workspace. Teams train agents, design journeys, test difficult conversations, launch outreach and supervise AI performance. Shared business records and organizational access connect these activities.

Executive Cockpit focuses on business impact and priorities. Platform Topology explains how business intent, AI, telephony and enterprise systems work together under human oversight. Agent Space brings customer context and AI assistance into the human agent experience.

Workspaces and value

Workspace	Purpose	Value to the operation
Executive Cockpit	Bring operational performance and business impact together	Help leaders prioritize investment and intervention
Agent Training	Practice conversations, assess readiness and coach	Build confidence and consistency before live customer work
Journey Studio	Design, test and version conversational behavior	Turn business requirements into controlled execution
Adversarial Testing	Challenge journeys with difficult calls	Find weak paths before broader exposure
Outbound Campaigns	Plan audiences, timing and authorized outreach	Coordinate customer contact and track outcomes
Agent Space	Combine customer context, assistance and resolution tasks	Help people resolve complex work; live assignment is part of the desktop roadmap
Fleet Control	Detect recurring issues and review improvements	Improve journeys through attributable, controlled experiments
Platform Topology	Connect services, dependencies and oversight	Give operators a shared understanding of the whole platform
Organization & User Management	Manage people, roles and configuration	Apply consistent access across shared business operations

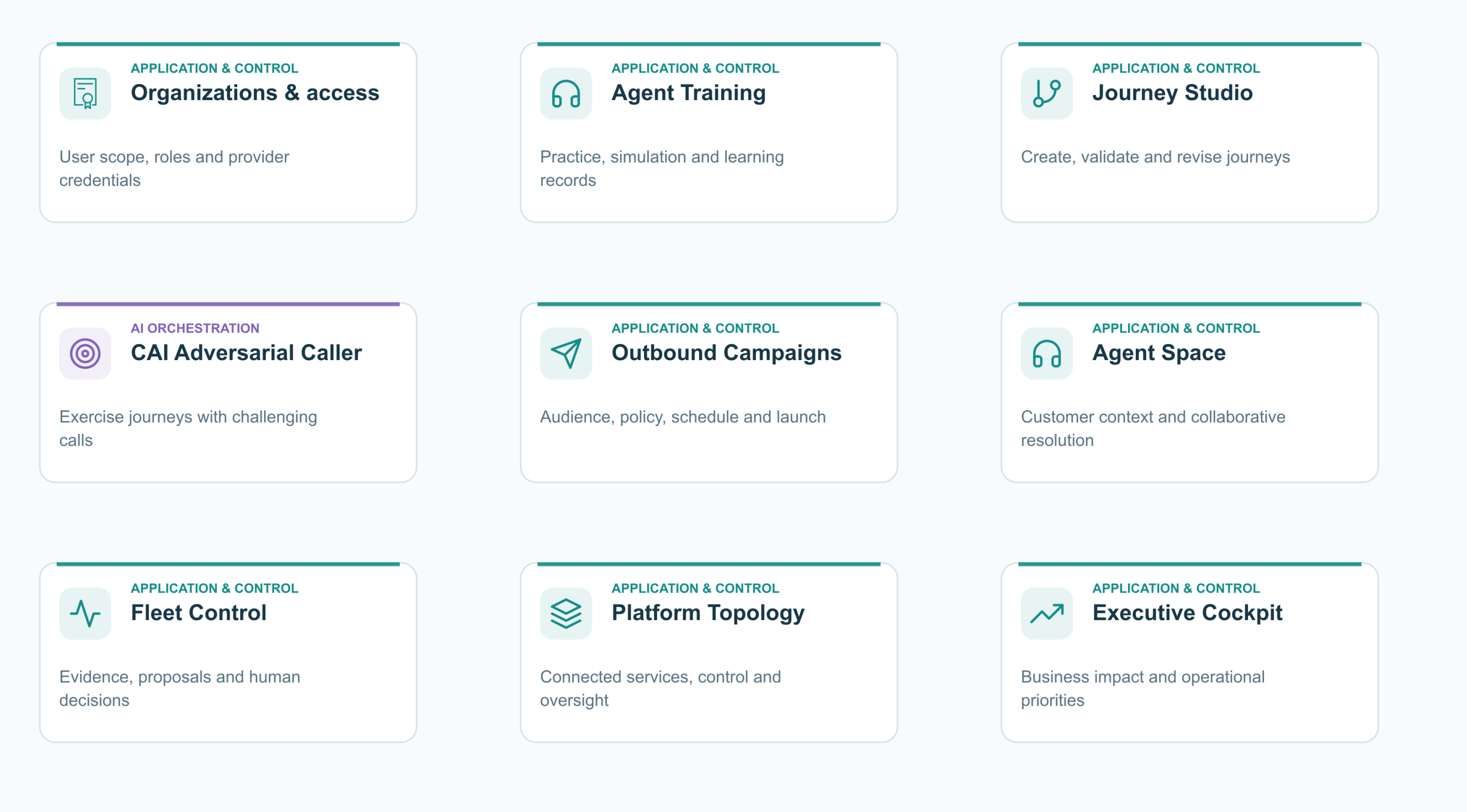
From insight to action

An unresolved topic can become a fleet issue, a reviewed journey improvement and a targeted coaching scenario. Teams carry context across the operating loop instead of recreating it in separate tools.

The roadmap connects Agent Space to durable queues and live sessions, and Platform Topology to service telemetry and approved infrastructure changes. Together these extend CAI into a complete operating environment for human and AI collaboration.

D17 · One platform, multiple workspaces

How is the experience organized?



The Command Center connects purposeful workspaces across the customer operation. Leaders prioritize business impact, teams develop agent readiness and designers refine journeys. Outbound Campaigns coordinates outreach, Fleet Control supports continuous improvement and Platform Topology connects the operating services. Agent Space brings customer context and AI assistance into the human resolution experience.

Roadmap & full CAI operation

CAI builds on a production voice foundation in two complementary directions: extend existing contact-center ecosystems, and deliver a complete CAI operating environment. Both use the same governed journeys, speech intelligence, telephony connectivity and business records.

The roadmap expands connection coverage, adds integrated queues and a live collaborative desktop, and extends supervised optimization from journeys to infrastructure. Each stage gives teams greater control over the customer experience.

Capability roadmap

Stage	Foundation	Next capability and value
Production voice	End-to-end PBX, SIP bridge, runtime and speech; in-house WebRTC and adversarial calling	Connect AI conversations to existing customer operations
Broader connectivity	SIP, provider media, WebSocket and WebRTC interfaces	Expand compatible clients, recovery options and deployment repeatability
Integrated customer operations	Agent Space, governed runtime and human control	Add queues, presence, assignment and a complete work lifecycle
Supervised optimization	Fleet observations, reviewed proposals and experiments	Extend improvement programs and approved infrastructure changes

Integrated queue and agent desktop

A durable work item connects intake, routing, human acceptance, AI assistance and follow-up. Skills, priority and available capacity determine assignment. Reservations prevent competing ownership; timeouts release or requeue work. Agent Space connects the assigned work to customer context and authorized session control.

Agents can collaborate with AI, take over a conversation or transfer a call according to the situation. Wrap-up records the outcome and creates attributable follow-up where needed. This lifecycle supports callbacks, overflow and supervisor intervention as the operation grows.

Supervised topology control

Platform Topology provides a shared view of the connected operation. The next stage links that view to service inventory, health and capacity telemetry. Proposed changes carry their scope, expected impact and approval owner. Authorized execution includes outcome verification and rollback.

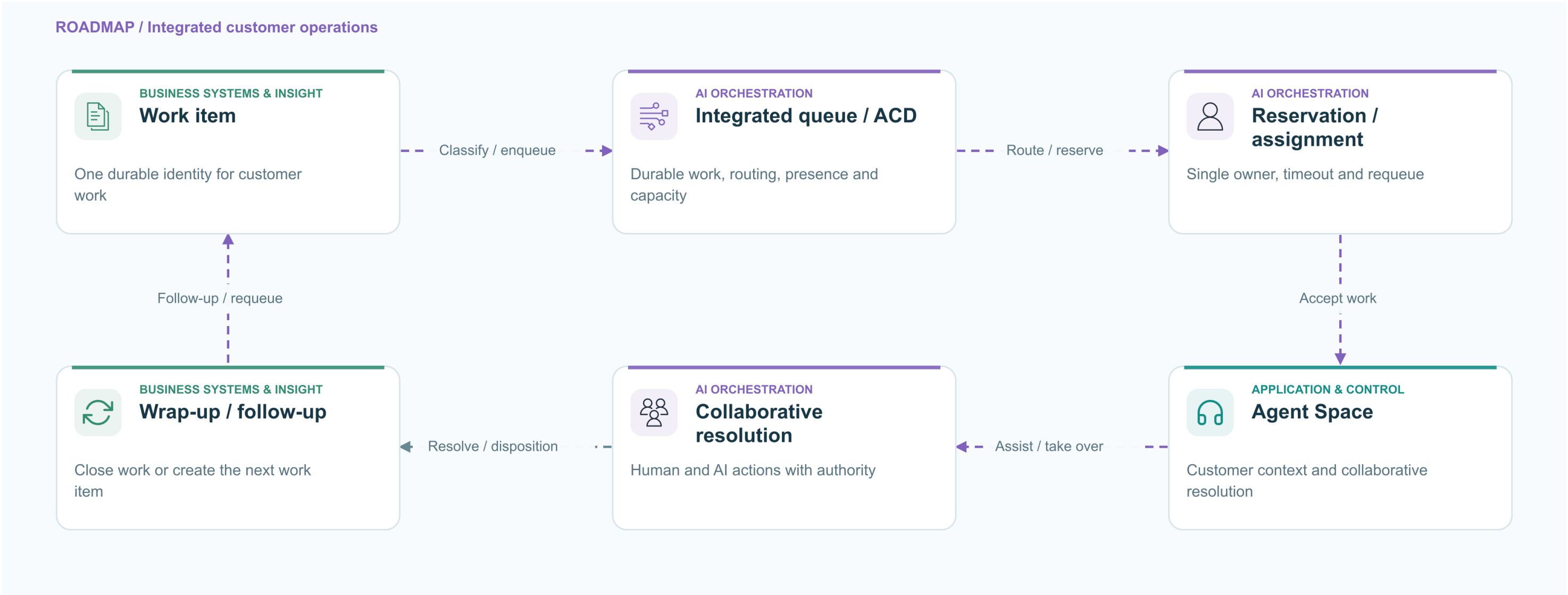
This extends the human control already used for journeys and fleet improvement to the infrastructure that runs them.

Two ways to adopt CAI

- Extend an existing contact center.** Retain the customer ACD, desktop, carrier and enterprise systems while adding CAI conversations, training, outbound execution and fleet improvement.
- Run the complete customer operation.** Combine those capabilities with integrated work distribution, agent presence, a collaborative desktop and supervised platform control.

D18 · Future queue and agent desktop

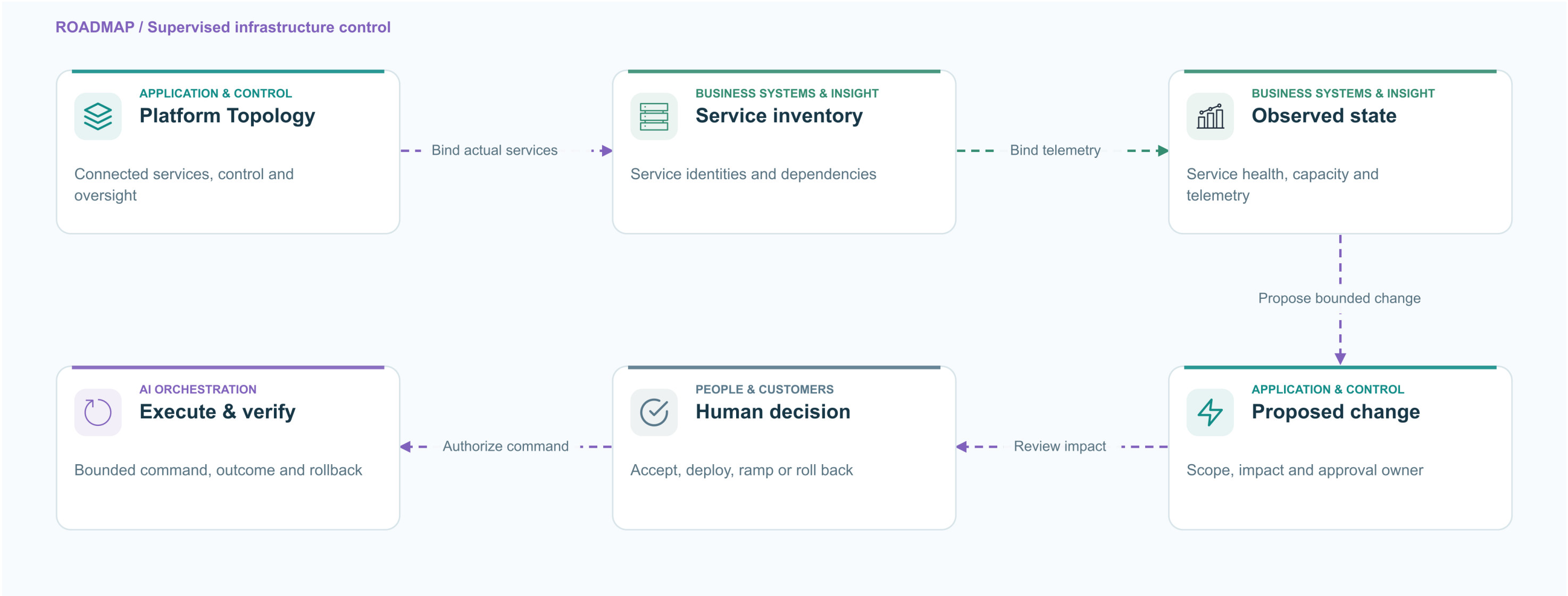
How does CAI own human work?



The integrated queue and desktop roadmap connects intake, routing, acceptance, resolution and follow-up through one durable work identity. Skills and capacity guide assignment; reservation prevents competing ownership; timeouts support requeueing. Agent Space connects the assigned work to customer context, AI assistance and authorized runtime control.

D19 · Supervised topology control

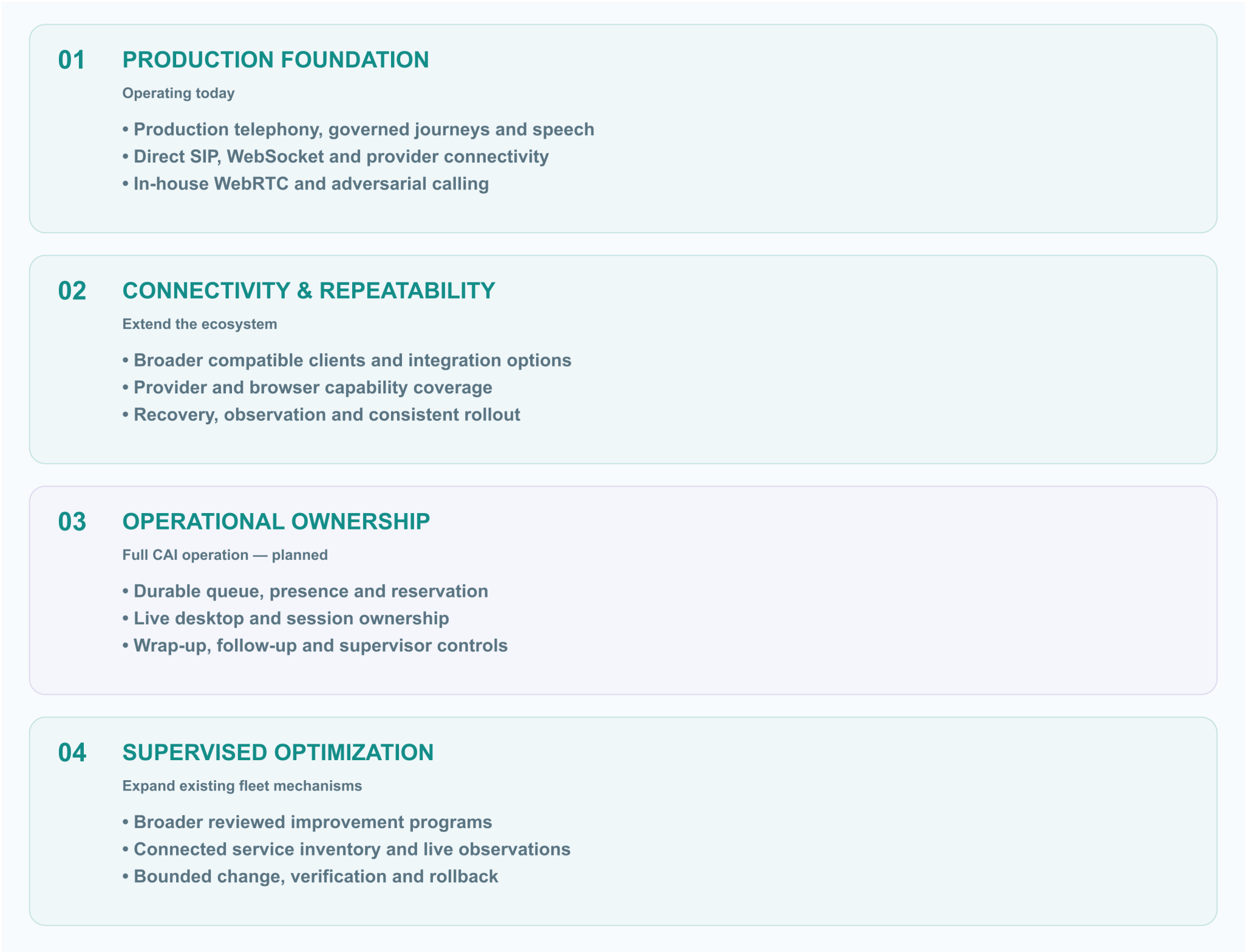
How will operators govern infrastructure changes?



The topology roadmap connects service inventory and live observations to supervised infrastructure changes. Each proposal carries its scope, expected impact and approval owner. Authorized execution verifies the outcome and supports rollback. Operators retain control while the system helps identify capacity, reliability and configuration improvements across connected services.

D20 · Roadmap from production foundation

What does each extension unlock?



CAI starts from a production voice foundation. Broader connectivity expands the ways customers integrate. Integrated queues and a live collaborative desktop extend CAI into a complete operating environment. Supervised optimization connects observed performance to reviewed changes in journeys and infrastructure. Both adoption paths reuse the same core platform and human authority.

Capabilities & terminology

CAI combines production voice connectivity, governed journey execution, speech intelligence and human supervision. Each capability contributes to a connected customer operation, from the first interaction to training and improvement.

The roadmap extends that foundation with integrated work distribution, a live collaborative desktop and supervised infrastructure control.

Connected operations

Business teams define objectives and journeys. CAI Agentic Runtime conducts the conversation and authorizes actions. CAI Speech Platform supplies voice intelligence. Interaction records connect execution to reporting, coaching and fleet improvement.

Capabilities

Production voice connectivity

In production

Connect existing PBX telephony through CAI SIP Media Bridge, CAI Agentic Runtime and CAI Speech Platform.

Governed execution

Available

Turn-taking, journey state, authorized tools, DTMF and human approval.

Interchangeable speech engines

Available

Native speech-to-speech, composed STT / LLM / TTS and hybrid engines behind one interface.

Browser voice

Available

In-house WebRTC within CAI Speech Platform.

Adversarial calling

Available

Repeatable challenging calls that exercise journeys and exception paths.

Fleet improvement

Available

Detect issues, review proposed changes and evaluate controlled journey experiments.

Tools and MCP

Supported

Connect business APIs and MCP tools through governed execution and selected results.

Integrated agent desktop

Roadmap

Connect Agent Space to live queue assignment, customer context and authorized session control.

Integrated queue / ACD

Roadmap

Durable work, skills, presence, reservation and recovery for full CAI operations.

Topology change control

Roadmap

Connect service inventory and telemetry to approved infrastructure changes, verification and rollback.

Container-based resilience

Deployment strategy

Docker packaging, redundant instances, health-based routing and controlled updates.

Terminology

Journey — Configured conversational behavior; the runtime commits its progression.

Revision — Immutable runnable journey snapshot. A ref is a movable pointer to it.

Deployment binding — A channel entry point bound to a journey revision/ref/strategy; distinct from infrastructure placement.

Outbound campaign — A planned run of contact attempts. Distinct from a fleet improvement campaign.

Fleet campaign — A bounded improvement run owning issues, interventions and experiments.

Twiml — HTTP-delivered telephony instructions; distinct from the compatible media WebSocket events.

SIP / RTP — SIP signals calls; RTP transports media. Neither is the runtime speech protocol.

ESL — FreeSWITCH Event Socket Library interface used for call-control commands.

S2S / STT / TTS — Speech-to-speech / speech-to-text / text-to-speech.

VAD — Voice activity detection: identifies speech boundaries for supported input paths.

MCP — Model Context Protocol: a standard interface for connecting AI conversations to tools and external systems.

WebRTC / TURN — Browser realtime media transport / optional network relay. WebRTC is an interface within CAI Speech Platform.