

Manual de referencia: Retell AI

Datos de esta manual

Fase	MAN — Manuales de referencia
Módulo	MAN — Biblioteca de consulta

Objetivo práctico

Este manual reúne, en un único documento de consulta, la arquitectura completa del panel de Retell AI (agentes, Retell LLM, voz, Custom Functions, Webhooks de cuenta, números, logs), sus planes y precios vigentes, la seguridad de sus claves de API, un examen honesto de la accesibilidad real de su panel con NVDA —incluidas las partes que presentan barreras— y el catálogo de errores frecuentes con su diagnóstico. Úsalo como referencia puntual, no como lectura lineal.

Resultado que obtendrás

Un documento de consulta permanente al que volver mientras trabajas con Retell AI: desde la primera configuración de un agente hasta el diagnóstico de un fallo concreto en producción meses después.

Dónde encaja dentro del sistema final

Dentro del sistema TELÉFONO → AGENTE → CEREBRO → VOZ → AUTOMATIZACIÓN → CALENDARIO, Retell AI es la segunda plataforma de referencia de este curso (alternativa a Vapi) para la pieza AGENTE: conecta el modelo de respuesta (Retell LLM, que hace de puente hacia el CEREBRO elegido), la voz (TTS) y, al asociarle un número, la puerta de entrada TELÉFONO. Sus Custom Functions y sus Webhooks de cuenta son el punto de conexión hacia AUTOMATIZACIÓN (n8n, Make) y, por extensión, hacia CALENDARIO.

Conocimientos previos

- Haber completado la lección f2-m3-l2-retell.php (ensamblaje del agente en Retell AI) dentro de la progresión del curso.
- Prompt Maestro completo (módulo 1.3).
- Concepto de Webhook y Tool Calling (módulos 1.2 y 3.4).
- Protocolo de creación segura de cuentas y de gestión de claves de API del módulo 2.1.

Herramientas necesarias

- Una cuenta de Retell AI.
- Un navegador con NVDA activo.
- Un gestor de contraseñas para las claves de API.

Posible coste

Modelo de precios (verificado el 30 de agosto de 2026 contra la página oficial de precios, retellai.com/pricing): Retell factura por uso puro, sin cuota mensual base, mediante tres componentes independientes que se suman para obtener el coste por minuto: infraestructura de voz (aproximadamente 0,055 USD/min), texto a voz (entre 0,015 y 0,040 USD/min, según el proveedor de TTS elegido) y modelo de lenguaje (LLM), con un rango muy amplio, desde aproximadamente 0,003 USD/min en los modelos más económicos hasta cifras muy superiores en los modelos de alta gama. A esos tres componentes se suma el coste de telefonía (aproximadamente 0,015 USD/min, variable según el país). El coste combinado observado en configuraciones completas suele situarse aproximadamente entre 0,13 y 0,31 USD por minuto.

Funciones adicionales de producción con coste propio: base de conocimiento (0,005 USD/min), reducción de ruido avanzada (0,005 USD/min), eliminación de datos personales (PII, 0,01 USD/min), y control de calidad por IA (0,10 USD/min tras los primeros 100 minutos gratuitos).

Crédito y concurrencia gratuitos: Retell ofrece 10 USD de crédito gratuito para empezar a probar sin compromiso inicial, y 20 llamadas concurrentes gratuitas por cuenta.

Plan Enterprise: precio a medida, para volumen alto y requisitos adicionales.

Límite de gasto nativo: según la documentación oficial de precios consultada, no se menciona ninguna función de límite de gasto configurable (hard limit) dentro de la propia plataforma. Esta situación coincide con lo ya verificado en la lección de límites de gasto del módulo 2.1: aplica medidas externas (saldo mínimo, tarjeta virtual con tope propio, alertas bancarias, monitorización activa) si vas a hacer pruebas o despliegues extensos.

Advertencia de seguridad

Claves de API: según la documentación oficial de Retell, las claves de API son credenciales de portador (bearer credentials) que autentican tanto las peticiones a la API REST y a los SDK oficiales, como la validación de los Webhooks entrantes

desde Retell. Se gestionan desde el panel, en «Workspace → Settings → API Keys» (o «Accounts → API Keys», según la versión de la interfaz). Un espacio de trabajo puede tener varias claves de API, todas con el mismo nivel de permisos: Retell no ofrece, según la documentación consultada, una restricción de alcance por clave equivalente a la de otras plataformas, por lo que cualquier clave generada tiene acceso completo a la cuenta.

Clave designada para Webhooks: Retell designa automáticamente una de tus claves de API como clave usada para la autenticación de Webhooks entrantes; ten en cuenta cuál es esa clave concreta al rotar credenciales, para no invalidar accidentalmente la validación de tus Webhooks en producción.

Buenas prácticas generales: nunca compartas una clave de API en un repositorio público ni la incluyas en código de lado del cliente; consévala en variables de entorno de tu servidor y rota cualquier clave (crear una nueva, migrar las integraciones, borrar la antigua) si sospechas que se ha expuesto. Dado que, a diferencia de otras plataformas, todas las claves de un espacio de trabajo comparten el mismo nivel de acceso, trata cualquier filtración de clave de Retell como una filtración de acceso completo a la cuenta, no como un incidente parcial.

Custom Functions y Webhooks: valida siempre, en tu propio servidor, que las peticiones POST que recibes en la URL configurada para una Custom Function o para los Webhooks de cuenta proceden realmente de Retell (por ejemplo, verificando la firma o cabecera de autenticación correspondiente), en lugar de asumir que cualquier petición a esa URL es legítima.

Modelo mental

Igual que con Vapi, piensa en Retell como en otra centralita telefónica completa, con sus propios mandos y disposición, pero cumpliendo exactamente la misma función: conectar cerebro, oído y voz en un único agente coherente, y avisar a otros departamentos (mediante Custom Functions y Webhooks) cuando ocurre algo relevante durante la llamada.

Explicación

Según la documentación oficial de Retell AI verificada el 30 de agosto de 2026, un **agente** se crea desde el panel («Agents → Create an Agent»), y admite dos modos de construcción: **Single prompt** (conversación libre a partir de un único prompt, el enfoque compatible con la estructura de Prompt Maestro de este curso) o **Conversational flow** (un flujo determinista de nodos conectados, más rígido y orientado a procesos muy estructurados, con ramificaciones de estado, nodos de

función y nodos de respuesta predefinida).

Retell LLM: el motor de respuesta se configura como una entidad independiente («Retell LLM») que se asocia al agente y contiene el prompt principal; esto permite, en teoría, reutilizar el mismo Retell LLM en varios agentes distintos.

Custom Functions: cuando el LLM decide invocar una función personalizada durante la conversación, Retell envía una petición POST a la URL que hayas configurado, con el nombre de la función y sus parámetros, exactamente el mismo mecanismo de Webhook estudiado en el módulo 1.2, aplicado aquí al Tool Calling del módulo 3.4.

Webhooks de cuenta: además de las Custom Functions asociadas a un agente concreto, Retell ofrece un sistema de Webhooks generales de cuenta, para recibir notificaciones de eventos del ciclo de vida de la llamada (inicio, fin, entre otros) sin necesidad de consultar la API repetidamente («polling»).

Conversational Flow — construcción visual: el modo «Conversational flow» se construye mediante un lienzo visual de nodos conectados con líneas («Function Nodes» para ejecutar acciones concretas, «Response Nodes» para respuestas predefinidas condicionadas a la entrada del usuario, entre otros tipos), disposición que se organiza principalmente arrastrando bloques sobre el lienzo.

Vocabulario nuevo

Custom Function (Retell)

Piénsalo así: El mismo recado que el telefonista envía al departamento correspondiente mediante el buzón (Webhook).

Función personalizada que el LLM de un agente Retell puede invocar durante la conversación, enviando una petición POST a una URL configurada por ti.

Ejemplo: Una función «reservar_cita» que envía los datos de la cita a un Webhook de n8n.

Retell LLM

Piénsalo así: El guion y el criterio de decisión del telefonista, separados del propio mueble de la centralita.

Entidad de configuración independiente dentro de Retell que contiene el prompt principal y el modelo de lenguaje elegido, y que se asocia a uno o varios agentes.

Ejemplo: Un mismo Retell LLM reutilizado en dos agentes que atienden números distintos del mismo negocio.

Conversational Flow

Piénsalo así: Un diagrama de flujo dibujado a mano, con cajas y flechas, en lugar de un guion escrito en prosa.

Modo de construcción de agentes de Retell basado en un lienzo visual de nodos conectados, orientado a procesos deterministas y muy estructurados, en contraste con el modo Single prompt.

Ejemplo: Un flujo de verificación de identidad con nodos de pregunta, condición y transferencia a un humano.

Webhook de cuenta (Retell)

Piénsalo así: Un aviso automático a la centralita general de la empresa cada vez que ocurre algo en cualquier llamada, no solo en una concreta.

Sistema de notificaciones de eventos del ciclo de vida de las llamadas a nivel de cuenta completa, alternativo a consultar la API repetidamente.

Ejemplo: Un Webhook de cuenta que registra en una hoja de cálculo cada llamada finalizada, con independencia del agente que la atendió.

Preparación

Ten a mano tu Prompt Maestro, tu clave de API del proveedor de voz elegido, y decide de antemano si vas a usar el modo Single prompt (recomendado por este curso) o Conversational flow antes de empezar a construir, porque cambiar de modo a mitad de proceso obliga a reconstruir la configuración desde cero.

Procedimiento paso a paso

1. **Contexto:** Vas a generar una clave de API para conectar tu propio backend con Retell.

Acción de teclado: Recorre encabezados con **H** hasta «Settings → API Keys» (o «Accounts → API Keys», según la versión). Activa el botón de creación de clave

con **Enter**.

Respuesta esperada de NVDA: NVDA debería anunciar un texto similar a «Create API Key, botón» y, tras activarlo, un cuadro de confirmación con la clave generada.

Qué significa: Recuerda que, a diferencia de Vapi, todas las claves de un mismo espacio de trabajo comparten el mismo nivel de permisos: no hay clave «restringida» de menor alcance.

Acción siguiente: Copia la clave a tu gestor de contraseñas de inmediato y confirma cuál de tus claves ha quedado designada por Retell para la autenticación de Webhooks.

2. **Contexto:** Vas a configurar una Custom Function conectada a un Webhook externo.

Acción de teclado: Dentro del editor del Retell LLM asociado a tu agente, recorre encabezados con **H** hasta la sección de funciones («Functions» o «Custom Functions»). Activa «Add Function» con **Enter**, completa con **Tab** el nombre, la descripción, los parámetros y la URL de destino.

Respuesta esperada de NVDA: NVDA anunciará cada campo del formulario conforme lo recorras, incluidos los campos de tipo de cada parámetro.

Qué significa: La descripción de la función es lo que el LLM lee para decidir cuándo invocarla: sé específico sobre en qué momento de la conversación debe usarse.

Acción siguiente: Guarda la función, pruébala mediante la opción de prueba del agente, y confirma en tu servidor que la petición POST llega con el formato esperado.

3. **Contexto:** Vas a revisar los logs de una llamada ya finalizada.

Acción de teclado: Recorre encabezados con **H** hasta «Call History» o «Logs». Localiza la llamada en la tabla de resultados con **Tab** o con la navegación por tablas de NVDA (**Ctrl** + **Alt** + flechas), y actívala con **Enter** para abrir el detalle.

Respuesta esperada de NVDA: NVDA debería anunciar las columnas de la tabla (fecha, duración, estado, agente) y, al abrir el detalle, la transcripción y los eventos registrados.

Qué significa: El detalle de la llamada es la fuente más fiable para diagnosticar

un comportamiento inesperado del agente.

Acción siguiente: Descarga la grabación o la transcripción si necesitas revisarla fuera del panel.

Posibles diferencias de interfaz

Retell distingue entre «Single prompt» y «Conversational flow» como dos formas distintas de construir un agente; este curso usa siempre «Single prompt» por su compatibilidad directa con la estructura de Prompt Maestro de la Fase 1 y por ser el modo accesible con NVDA (ver más abajo). Si tu versión del panel presenta estas opciones con nombres ligeramente distintos, busca por los conceptos («prompt único» frente a «flujo por pasos»).

Problema de accesibilidad y alternativa

Barrera real identificada: el modo «Conversational Flow» se construye sobre un lienzo visual de nodos conectados mediante interacción de arrastre (drag and drop): posicionar un nodo, conectarlo a otro mediante una línea, y organizar la ramificación del flujo dependen de manipular elementos gráficos sobre un canvas. Esta interacción no tiene, según la documentación pública consultada, un equivalente declarado de teclado o de lector de pantalla dentro del propio editor visual, lo que constituye una barrera de accesibilidad real para construir o modificar flujos complejos exclusivamente desde el panel visual con NVDA.

Alternativa disponible: el modo «Single prompt» —el que usa este curso, y el recomendado por esa misma razón de accesibilidad además de por su compatibilidad con el Prompt Maestro— se configura íntegramente mediante formularios estándar (campos de texto, selectores), sin ningún lienzo gráfico, y es completamente operable con NVDA. Si un proyecto concreto exigiera en el futuro un Conversational Flow determinista, la vía recomendada es definir la lógica del flujo mediante la API de Retell (que acepta la estructura completa de nodos y conexiones como datos, no como manipulación gráfica), en lugar de operar el lienzo visual directamente. Verifica siempre, antes de asumir esa vía como definitiva, que el endpoint de creación de flujos cubre en tu momento de consulta la misma expresividad que el editor visual, ya que esta es un área de la plataforma sujeta a cambios frecuentes.

Errores frecuentes y diagnóstico

No encuentras dónde pegar el Prompt Maestro completo.

Cómo localizarlo: En el modo «Conversational flow», el prompt se reparte en varios nodos en lugar de un único campo, lo cual no es compatible con la estructura de bloque único del Prompt Maestro de este curso.

Solución: Confirma que has elegido el modo «Single prompt» al crear el agente; si ya lo creaste con el modo equivocado, créalo de nuevo con el modo correcto en lugar de intentar convertirlo.

Una Custom Function nunca recibe la petición esperada en el servidor.

Cómo localizarlo: Revisa que la URL configurada en la Custom Function sea correcta y esté accesible, y consulta el detalle de logs de la llamada para confirmar si Retell intentó la petición y qué respuesta obtuvo.

Solución: Corrige la URL o la descripción de la función, y usa un endpoint de prueba temporal para confirmar el formato exacto de la petición recibida.

Los Webhooks de cuenta dejan de autenticarse correctamente tras rotar una clave de API.

Cómo localizarlo: Retell designa automáticamente una de tus claves para autenticar los Webhooks; si esa clave concreta fue la que rotaste o borraste, la validación deja de superarse en tu servidor.

Solución: Revisa en el panel qué clave está designada actualmente para Webhooks antes de rotar credenciales, y actualiza tu servidor con la nueva clave si cambia.

El coste real de las llamadas varía mucho entre pruebas, aunque la voz y la telefonía sean las mismas.

Cómo localizarlo: El coste de Retell se compone de varios elementos independientes (infraestructura de voz, TTS, LLM y telefonía), y el precio del componente LLM tiene un rango muy amplio según el modelo elegido (desde modelos económicos hasta modelos de alta gama).

Solución: Revisa qué modelo de LLM está asociado al Retell LLM de cada agente antes de comparar costes entre pruebas, y documenta la combinación exacta de proveedores usada en cada presupuesto.

Comprobación final

Confirma que puedes, usando solo NVDA: crear una Custom Function con su URL de destino, localizar el detalle de logs de una llamada de prueba anterior, y explicar por qué el modo Single prompt es preferible a Conversational Flow cuando trabajas exclusivamente desde el panel visual. Si consigues las tres cosas, la comprobación es positiva.

Qué acabas de conseguir

Tienes una referencia completa de la arquitectura de Retell AI —agentes, Retell LLM, Custom Functions, Webhooks de cuenta y logs—, con un criterio claro sobre qué partes del panel son accesibles de forma directa con NVDA y cuáles requieren la API como alternativa.

Ejercicio práctico

Compara, con el mismo Prompt Maestro y las mismas preguntas de prueba, el coste estimado por minuto de tu agente en Retell frente al de tu asistente en Vapi (manual de referencia correspondiente), documentando qué componentes concretos explican la diferencia.

Resumen de lo imprescindible

- Retell factura por tres componentes independientes (infraestructura de voz, TTS y LLM) más telefonía, sin cuota mensual base; el coste combinado observado suele situarse entre 0,13 y 0,31 USD/min.
- Todas las claves de API de un espacio de trabajo de Retell comparten el mismo nivel de permisos: trata cualquier filtración como un acceso completo a la cuenta.
- El modo Single prompt es compatible con el Prompt Maestro de este curso y completamente accesible con NVDA; Conversational Flow usa un lienzo visual de arrastre que constituye una barrera real de accesibilidad.
- Las Custom Functions y los Webhooks de cuenta son el mecanismo de conexión con AUTOMATIZACIÓN, mediante peticiones POST a una URL que debes validar en tu propio servidor.
- La documentación oficial de precios de Retell no menciona un límite de gasto configurable nativo: aplica medidas externas de control.

Estoy preparado para continuar si puedo...

- Explicar la diferencia entre Single prompt y Conversational Flow en Retell, y cuál es accesible con NVDA.
- Configurar una Custom Function con su URL de destino y diagnosticar por qué no llega una petición esperada.
- Calcular el coste real por minuto de una configuración de Retell, sumando sus tres componentes independientes.
- Explicar por qué todas las claves de API de Retell deben tratarse como de acceso completo a la cuenta.

Fuentes

- [Retell AI Pricing — Retell AI \(consultado el 2026-08-30\)](#)
- [Retell AI — Custom Function — Retell AI \(consultado el 2026-08-30\)](#)
- [API Key Overview — Retell AI \(consultado el 2026-08-30\)](#)
- [Retell AI — build, test, deploy, and monitor voice & chat agents \(portal de documentación\) — Retell AI \(consultado el 2026-08-30\)](#)
- [Retell AI's Advanced Conversation Flow — Retell AI \(consultado el 2026-08-30\)](#)

Última actualización de este contenido: **2026-08-30**.