Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Ciudad de México

Inteligencia Artificial Avanzada para la Ciencia de Datos II

TC3007C.500

Reporte Técnico

Autores

Guillermo Ian Barbosa Martínez - A01747926
Wilfrido Tovar Andrade - A01664769
Arely Yael Villatoro Amador - A01663303
Ariel López García - A01275913
Carlo Crivelli Hernández - A01656171
Ian Gabriel Rivera Reyes - A01658613

Profesores

David Christopher Balderas Silva

Jesús Manuel Vázquez Nicolás

Emmanuel Páez López

José Ángel Martínez Navarro

Óscar Francisco Fuentes Casarrubias

Grupo 100

1 de noviembre de 2025

Índice

Introducción

Problemática

Cada mes en los centros de reparación los técnicos tienen que enfrentarse al lanzamiento de productos nuevos, la amplia variedad de productos y la alta demanda de reparaciones diarias no les permite conocer a fondo todos los dispositivos con los que trabajan. Al revisar videotutoriales y foros tienen que detener su trabajo por periodos de tiempo considerables, y cuando utilizan chat GPT para solucionar sus dudas se encuentran con el desafío de que han pasado meses desde su última actualización por lo que no cuenta con la información más reciente y en muchas ocasiones brinda información errónea o incompleta. Esto genera retrasos que afectan el tiempo de entrega que viene ligada con la satisfacción de los clientes.

Los técnicos de campo requieren de una herramienta que les permita acceder a información clara y puntual de los dispositivos con los que están trabajando, sin necesidad de detener su trabajo.

Objetivo

Para optimizar la búsqueda de información de los técnicos de campo se creará un asistente, con base en un LLM, especializado en los dispositivos con los que trabajan y actualizado para brindar información clara, puntual y posibles soluciones acorde al dispositivo y modelo solicitado.

Alcance de la Solución

El presente proyecto abarca el diseño, desarrollo e implementación de un asistente digital basado en inteligencia artificial para soporte técnico de campo, enfocado exclusivamente en la explotación y consulta de manuales técnicos almacenados en repositorios controlados (*Oracle Object Storage*). El sistema incluirá una canalización automatizada de ingestión de documentos (subida, extracción de texto y segmentación en *chunks*), el almacenamiento estructurado de los fragmentos resultantes en una base de datos autónoma, y la capa de recuperación (*RAG*) que recupera y entrega fragmentos relevantes al usuario. Se integrará una interfaz web construida en *Oracle APEX* que permita a los técnicos realizar búsquedas por texto natural, consultar el contenido de manuales y recibir respuestas generadas por un *LLM* gobernado por reglas (*guardrails*): el asistente debe referenciar únicamente información extraída de los manuales cargados y señalar explícitamente cuando una consulta esté fuera del dominio documental.

Dentro del alcance funcional se incluyen:

1. Mecanismos de ingestión y procesamiento de PDFs (*chunking*, página por página con límites de tokens).

2. Carga y versionado de los metadatos y fragmentos en tablas relacionales diseñadas para RAG y búsqueda por texto.
3. Implementación de dos fuentes de *retrieval augmented generation*.
4. Desarrollo de la interfaz APEX con campo de entrada, botones de búsqueda, presentación de resultados y página de detalle por manual.
5. Configuración de PARs necesarios para que el servicio sea seguro y replicable en el entorno de *OCI Free Tier*.

Quedan fuera del alcance de esta fase las siguientes capacidades: diagnóstico visual (análisis automático de imágenes o fotos del equipo), ejecución remota de acciones en dispositivos, integración con sistemas externos de *ticketing* o ERP (salvo exportación manual de reportes), y generación de instrucciones que no estén respaldadas por el contenido de los manuales. A su vez, la sincronización masiva y la indexación incremental serán tratadas como mejoras posteriores.

Como resultados concretos del proyecto se entregarán:

a. El *bucket* configurado con PARs y el proceso de ingestión automatizado.
b. La(s) tabla(s) en la Base de Datos Autónoma que contengan metadatos y fragmentos de texto (con su esquema documentado).
c. Scripts/notebooks reproducibles para la extracción, *chunking*.
d. La aplicación APEX desplegada con las páginas de búsqueda y visualización de manuales
e. La configuración del motor de RAG y del *prompt system* (incluyendo guardrails)
f. Un conjunto de pruebas de evaluación con métricas cuantitativas que permitan validar precisión y relevancia en la recuperación de información.

Se consideran supuestos operativos: que los manuales suministrados están en formatos legibles (PDF con texto o escaneos con OCR factible), que la cuenta OCI dispone de los permisos y cuotas mínimas para *Object Storage*, *Generative AI* y *Autonomous Database*, y que el equipo técnico dispone de las credenciales necesarias para crear usuarios/roles en la base de datos y configurar APEX. Los criterios de aceptación incluyen: la capacidad del asistente para recuperar y citar fragmentos de manual relevantes ante consultas reales de técnicos.

Roles

- Guillermo
    - Desarrollador
    - Tester
    - Documentador
- Wilfrido
    - Desarrollador
    - Tester
    - Documentador

- Arely
    - Desarrolladora
- Ariel
    - Desarrollador
    - Tester
    - Documentador
- Ian
    - Desarrollador
    - Tester
- Carlo
    - Desarrollador

Planificación de tareas

El diagrama de Gantt con las tareas asignadas a cada miembro del equipo se encuentra en la siguiente liga: https://tecmx-my.sharepoint.com/:x:/g/personal/a01664769_tec_mx/Ef2qIjtCzxlGlCRwRC4hFAcBvQcHEWj320LsQPLZ4TlG4Q?e=eFIJp4

Usar las flechas de la barra de navegación para moverse entre las fechas

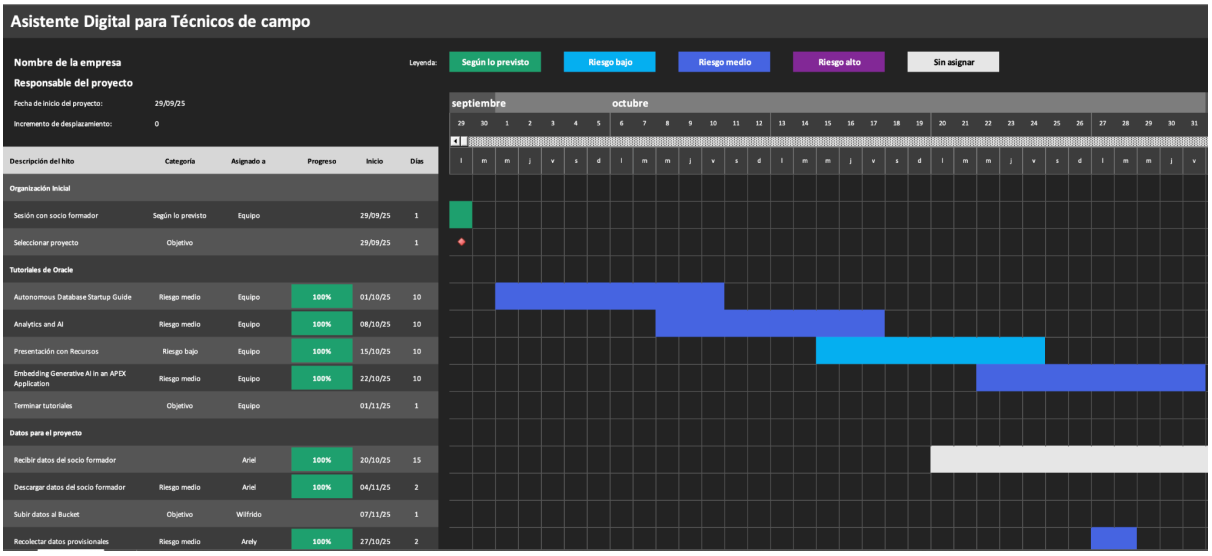
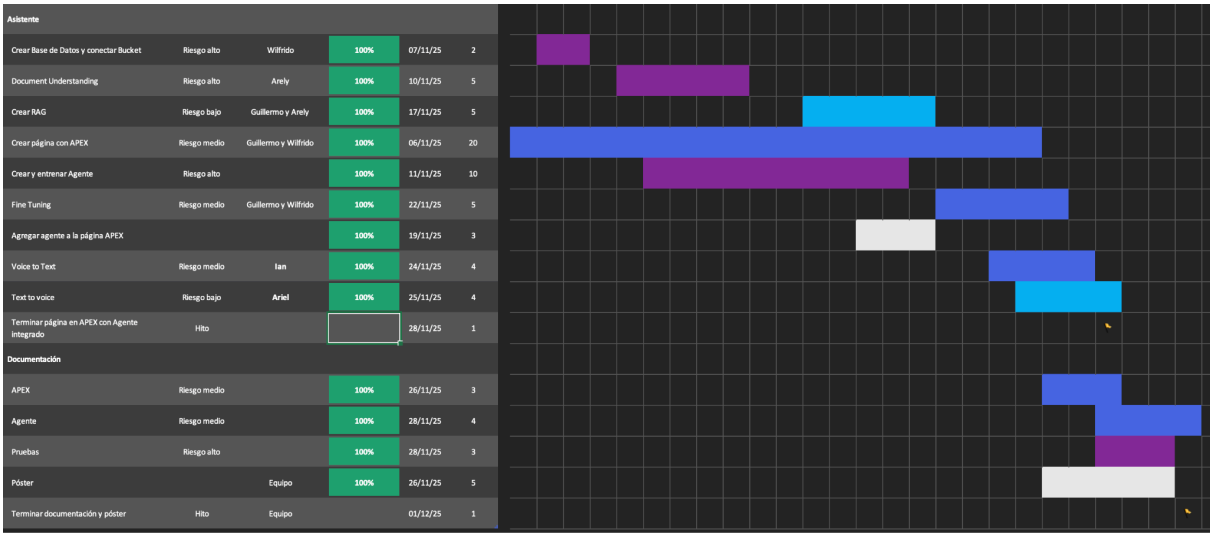


Figura 1. Planificación de actividades iniciales para la realización del proyecto.



Asistente					
Crear Base de Datos y conectar Bucket	Riesgo alto	Wilfrido	100%	07/11/25	2
Document Understanding	Riesgo alto	Arely	100%	10/11/25	5
Crear RAG	Riesgo bajo	Guillermo y Arely	100%	17/11/25	5
Crear página con APEX	Riesgo medio	Guillermo y Wilfrido	100%	06/11/25	20
Crear y entrenar Agente	Riesgo alto		100%	11/11/25	10
Fine Tuning	Riesgo medio	Guillermo y Wilfrido	100%	22/11/25	5
Agregar agente a la página APEX			100%	19/11/25	3
Voice to Text	Riesgo medio	Ian	100%	24/11/25	4
Text to voice	Riesgo bajo	Ariel	100%	25/11/25	4
Terminar página en APEX con Agente integrado	Hito			28/11/25	1
Documentación					
APEX	Riesgo medio		100%	26/11/25	3
Agente	Riesgo medio		100%	28/11/25	4
Pruebas	Riesgo alto		100%	28/11/25	3
Póster		Equipo	100%	26/11/25	5
Terminar documentación y póster	Hito	Equipo		01/12/25	1

Figura 2. Planificación de actividades finales para la realización del proyecto.

Exploración del proyecto

Técnicos de campo

Son personas especializadas en la instalación, programación, mantenimiento y reparación de diferentes equipos electrónicos que se encuentran en plantas o lugares similares.
Al interactuar directamente con los clientes cargan con una gran responsabilidad para mantener la reputación de la empresa a la que representan, requieren de acceso directo y rápido a la información necesaria para arreglar el equipo al igual que para responder las dudas que el cliente pregunte en el momento para asegurar su satisfacción
Cuándo técnico enviado no cuenta con los recursos y la experiencia para volver a poner las cosas en marcha, no solo genera dolores de cabeza para los clientes, sino que también afecta a los técnicos con un trabajo que podría tener horarios largos e impredecibles disminuyendo su rendimiento y retrasando otras tareas que tienen asignadas.
De acuerdo con Rodriguez (2024), la IA y las tecnologías basadas en datos pueden trabajar junto con los técnicos de primera línea para acelerar las respuestas y echar una mano cuando están en el campo. Estos sistemas inteligentes ayudan a ahorrar tiempo y complementan los conocimientos de su equipo para una experiencia más gratificante en el trabajo.

Herramientas usadas

Para llevar a cabo nuestro proyecto decidimos usar las herramientas de Oracle porque incluye las herramientas necesarias para almacenar nuestros manuales y procesarlos con diferentes herramientas integradas con Inteligencia Artificial, al igual que cuenta con una plataforma low code para crear la aplicación web donde desplegamos nuestro asistente.

Las herramientas de Oracle que vamos a implementar son las siguientes:

Document Understanding

Este servicio impulsado por IA puede extraer texto y sus puntos clave de cualquier tipo de documento, tiene su propia API para conectarlo a herramientas externas de Oracle. Usaremos esta herramienta para automatizar la extracción de la información. Lo ajustaremos para que funcione con los manuales y con la información extraída entrenaremos nuestro modelo.

Speech Recognition

Incluye la funcionalidad doble para transcribir audios de voz en tiempo real y sintetiza voz a partir de texto. Contiene varios modelos para trabajar con diferentes lenguajes y entender conversaciones naturales; al sintetizar la voz utiliza IA para crear diferentes tipos de voz que suenen naturalmente.

Text to voice

Este servicio es brindado mediante dos partes, la primera es OpenAI con su servicio de TTS desde su modelo GPT-4o-mini-TTS que es barato, eficiente y con capacidad de brindar su prestación mediante API y en caso de no contar con suficientes créditos o por algún motivo no se puede acceder entonces se utiliza el speech predeterminado de cada navegador; fue diseñado con la opción automática de reproducir los audios en cuanto se generen pero debido a que en ocasiones pueden estar bloqueadas estas configuraciones desde diversos dispositivos se optó por también incluir la opción de un botón para reproducir manualmente cada audio. Esto facilita enormemente el uso para cada técnico de campo, haciendo que pueda ser utilizado inclusive sin la necesidad de estar constantemente en contacto con sus dispositivos.

Voice to Text

Speech Recognition (Speech-to-Text)
Para este proyecto implementamos un sistema de reconocimiento de voz directamente en la aplicación web utilizando la Web Speech API, una tecnología nativa de los navegadores modernos que permite convertir audio en texto sin necesidad de servicios externos adicionales. Esta herramienta nos permite capturar la voz del técnico en campo y transformarla en texto de manera inmediata dentro de Oracle APEX.

Agentic AI

Dentro de los servicios de Oracle podemos crear agente impulsado con AI entrenados con la información que nosotros especifiquemos para que sea especialista en el tema deseado. Al igual que otras herramientas puede entender y responder en otros idiomas. Su principal ventaja es que se pueden establecer reglas y limitaciones a su comportamiento.

APEX

Esta plataforma nos permite crear aplicaciones web que puedan correr en cualquier dispositivo desde la nube o forma local. Gracias a su integración con la nube de Oracle se

pueden conectar otros servicios y aplicaciones. Las aplicaciones creadas con APEX se pueden desplegar fácilmente con Oracle Cloud y gracias a su diseño low code no se requiere de conocimientos en desarrollo web para crear aplicaciones.

Desarrollo del modelo

El proyecto se apoya en varios servicios de Oracle Cloud: los manuales se alojan en Object Storage y se comparten mediante *Pre-Authenticated Requests (PARs)* para permitir accesos controlados sin exponer credenciales; los PARs son objetos configurables desde la consola y la documentación oficial explica su uso y riesgos de seguridad.

ID	FILE_NAME	DESCRIPTION	FILE_URL
1	Amplificador de Señal 5G MaxBoost A-750	Manual Técnico	https://objectstorage.us-chicago
2	ONT (Terminal de Red Óptica) FT-1000	Manual Técnico	https://objectstorage.us-chicago
3	Router FiberOptix G-24	Manual Técnico	https://objectstorage.us-chicago
4	Access Point WiFi 6 AP-6000	Manual de Usuario	https://objectstorage.us-chicago
5	Access Point WiFi 6E AP-7800	Manual de Usuario	https://objectstorage.us-chicago
6	Amplificador RF Banda Dual BD-1200	Manual de Usuario	https://objectstorage.us-chicago
7	Analizador de Espectro SA-3000	Manual de Usuario	https://objectstorage.us-chicago
8	Antena Direccional Panel AD-2400	Manual de Usuario	https://objectstorage.us-chicago
9	Antena Sectorial AS-2100	Manual de Usuario	https://objectstorage.us-chicago
10	Balanceador de Carga LB-3200	Manual de Usuario	https://objectstorage.us-chicago
11	Central Telefónica IP PBX-8000	Manual de Usuario	https://objectstorage.us-chicago
12	Decodificador SmartTV STB-500	Manual de Usuario	https://objectstorage.us-chicago
13	Firewall Empresarial FW-8500	Manual de Usuario	https://objectstorage.us-chicago
14	Firewall de Red FW-5500	Manual de Usuario	https://objectstorage.us-chicago
15	Gateway VoIP SIP GW-4000	Manual de Usuario	https://objectstorage.us-chicago
16	IP PBX CentralVoice CV-8000	Manual de Usuario	https://objectstorage.us-chicago
17	Media Converter Fibra MC-1000	Manual de Usuario	https://objectstorage.us-chicago
18	Media Converter Gigabit MC-3000	Manual de Usuario	https://objectstorage.us-chicago
19	Modem Coaxial CableNet C-800	Manual de Usuario	https://objectstorage.us-chicago
20	Multiplexor DWDM MX-4800	Manual de Usuario	https://objectstorage.us-chicago
21	Multiplexor SDH_SONET MX-STM16	Manual de Usuario	https://objectstorage.us-chicago
22	Repetidor Celular DualBand RB-2600	Manual de Usuario	https://objectstorage.us-chicago
23	Repetidor Celular MultiCell RC-4800	Manual de Usuario	https://objectstorage.us-chicago
24	Router Empresarial RT-8500	Manual de Usuario	https://objectstorage.us-chicago
25	Router Satelital SR-4500	Manual de Usuario	https://objectstorage.us-chicago
26	Servidor de Base de Datos DB-9600	Manual de Usuario	https://objectstorage.us-chicago
27	Servidor de Video Streaming VS-8000	Manual de Usuario	https://objectstorage.us-chicago
28	Sistema UPS Industrial UPS-5000i	Manual de Usuario	https://objectstorage.us-chic

Figura 3. Acceso a manuales fuera del *bucket* con sus PARs. Esta tabla se configuró en la instancia de APEX correspondiente para dotar al modelo de acceso a los manuales.

La base de datos donde versionamos y servimos los fragmentos es una Autonomous Database configurada con *workload* de tipo *Transactional Processing* (diseñada para cargas OLTP/consultas rápidas y pequeñas transacciones), lo que optimiza la latencia y concurrencia de las consultas RAG desde APEX. Para extraer texto y metadatos de los PDFs usamos *Document Understanding* (*AI Service Document*) vía su API *analyzeDocument*, que permite enviar chunks de páginas y recibir líneas, claves y estructuras útiles para el RAG.

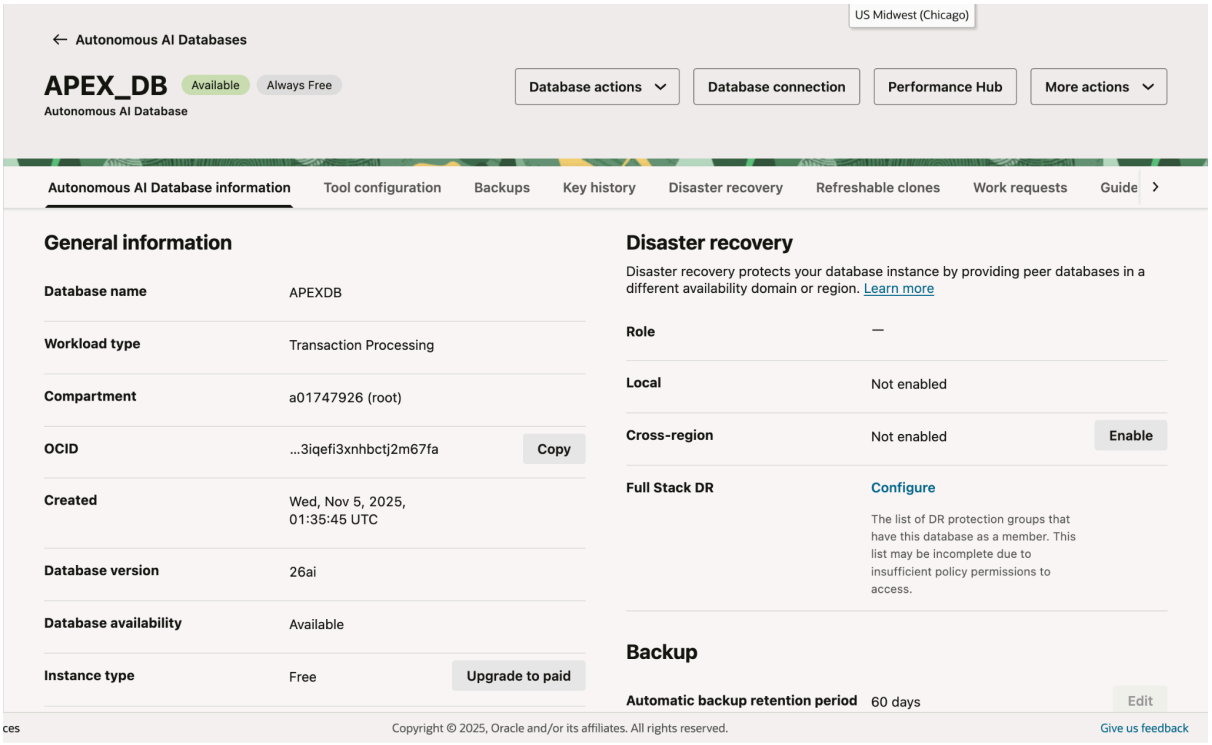
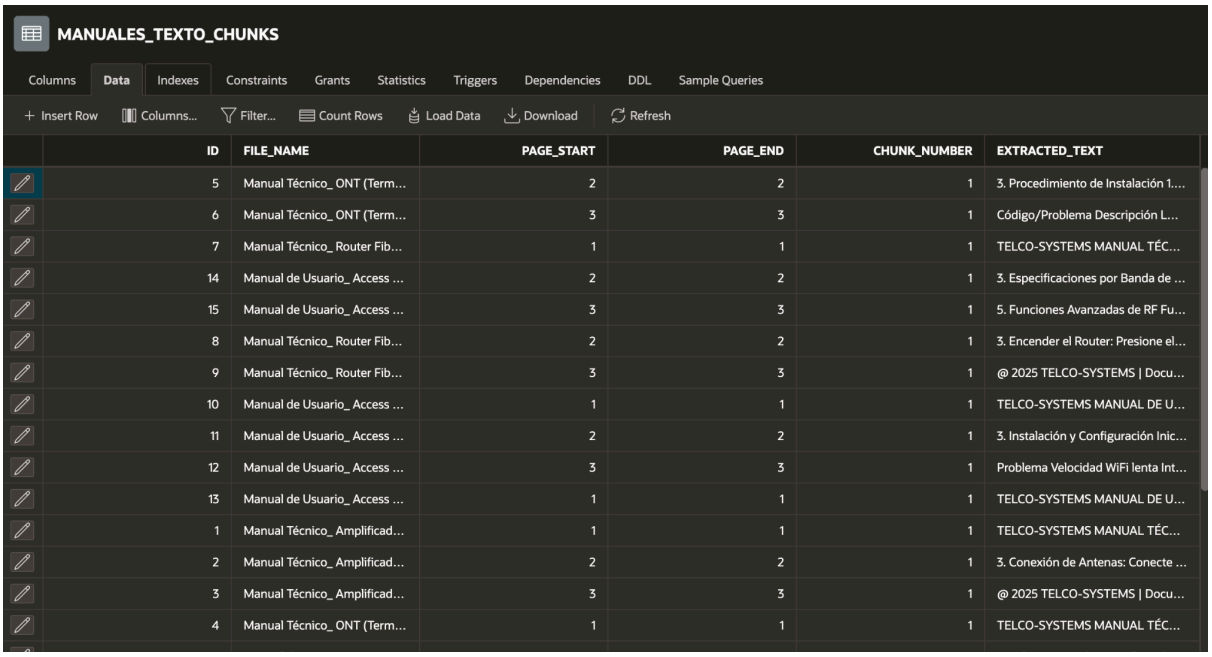


Figura 4. Base de datos autónoma desde la cual se lanzó la instancia de APEX.



ID	FILE_NAME	PAGE_START	PAGE_END	CHUNK_NUMBER	EXTRACTED_TEXT
5	Manual Técnico_ ONT (Term...	2	2	1	3. Procedimiento de Instalación 1....
6	Manual Técnico_ ONT (Term...	3	3	1	Código/Problema Descripción L...
7	Manual Técnico_ Router Fib...	1	1	1	TELCO-SYSTEMS MANUAL TÉC...
14	Manual de Usuario_ Access ...	2	2	1	3. Especificaciones por Banda de ...
15	Manual de Usuario_ Access ...	3	3	1	5. Funciones Avanzadas de RF Fu...
8	Manual Técnico_ Router Fib...	2	2	1	3. Encender el Router: Presione el...
9	Manual Técnico_ Router Fib...	3	3	1	@ 2025 TELCO-SYSTEMS | Docu...
10	Manual de Usuario_ Access ...	1	1	1	TELCO-SYSTEMS MANUAL DE U...
11	Manual de Usuario_ Access ...	2	2	1	3. Instalación y Configuración Inic...
12	Manual de Usuario_ Access ...	3	3	1	Problema Velocidad WiFi lenta Int...
13	Manual de Usuario_ Access ...	1	1	1	TELCO-SYSTEMS MANUAL DE U...
1	Manual Técnico_ Amplificad...	1	1	1	TELCO-SYSTEMS MANUAL TÉC...
2	Manual Técnico_ Amplificad...	2	2	1	3. Conexión de Antenas: Conecte ...
3	Manual Técnico_ Amplificad...	3	3	1	@ 2025 TELCO-SYSTEMS | Docu...
4	Manual Técnico_ ONT (Term...	1	1	1	TELCO-SYSTEMS MANUAL TÉC...

Figura 5. Carga de la información extraída, *chunk* por *chunk* y manual por manual, a la instancia de APEX.

El motor de respuestas lo ejecuta *OCI Generative AI* (configurado en APEX como servicio de Generative AI) usando un modelo instructivo de la familia LLaMA, específicamente el Meta Llama 4 Maverick-instruct de 17 mil millones de parámetros, un modelo orientado a tareas de razonamiento y comprensión técnica, que se registra en APEX como *AI Service* y se invoca desde las configuraciones de la app para generar respuestas condicionadas por el system

prompt y las fuentes RAG. Elegimos trabajar con el modelo de *Meta Llama 4 Maverick* porque representa el mejor equilibrio entre capacidad de razonamiento, contexto extremadamente amplio (hasta 512,000 tokens), arquitectura MoE eficiente con 128 expertos y soporte nativo para agentes, lo que lo hace ideal para un asistente técnico que debe analizar fragmentos extensos, citar con precisión y resolver instrucciones procedimentales. A diferencia de los modelos de *Cohere* —más orientados a *throughput* y *tool-use*— o los *OpenAI gpt-oss* —excelentes en STEM pero sin capacidades multimodales ni el mismo rango contextual— Maverick ofrece mayor robustez para tareas de recuperación y análisis detallado, manteniendo un desempeño superior en instrucciones complejas y en ambientes integrados como OCI APEX.

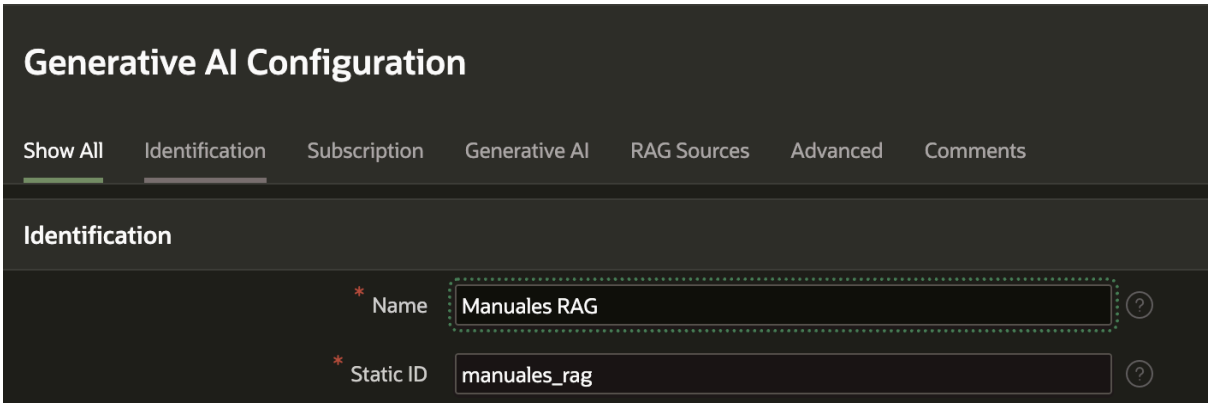


Figura 6. Primera etapa de la configuración del asistente de *Generative AI*.

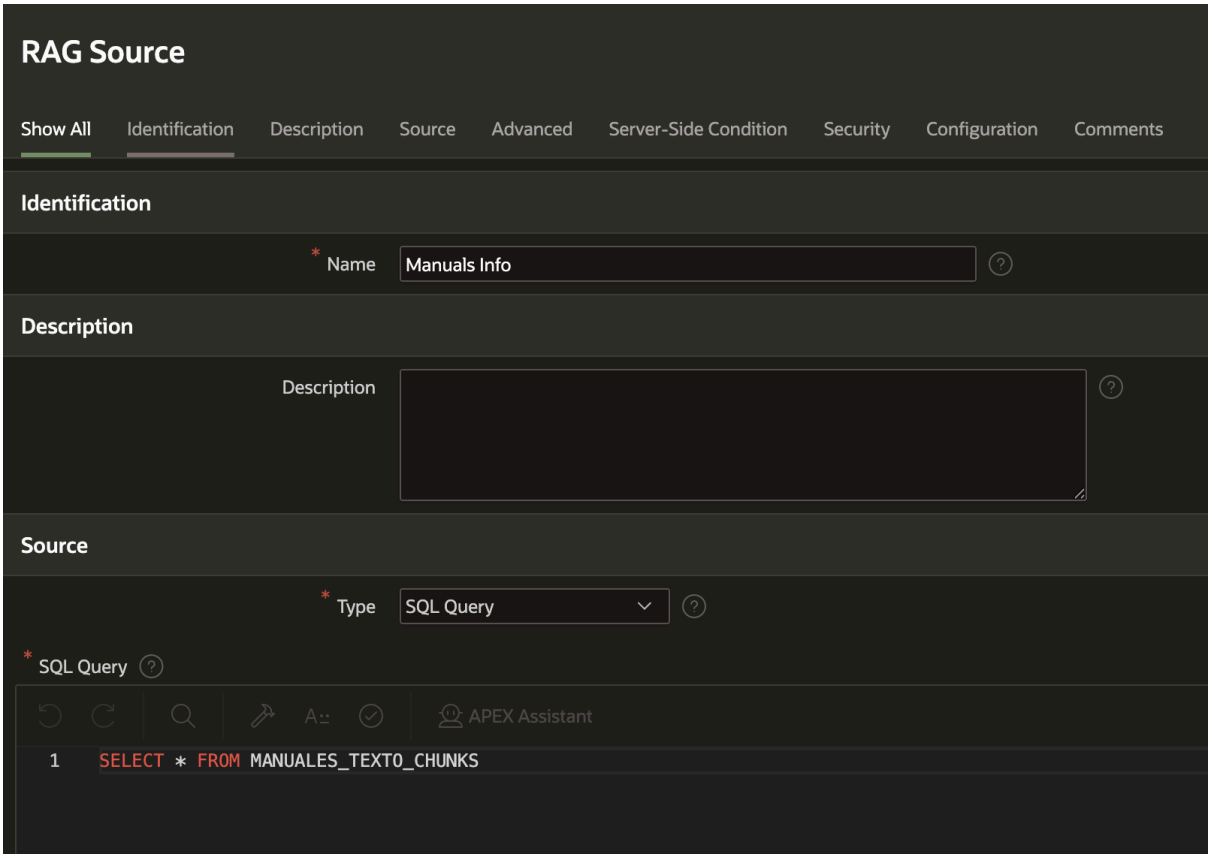


Figura 7. Primer RAG Source, que permite al modelo acceder a toda la información disponible en los manuales, *chunk* por *chunk*.

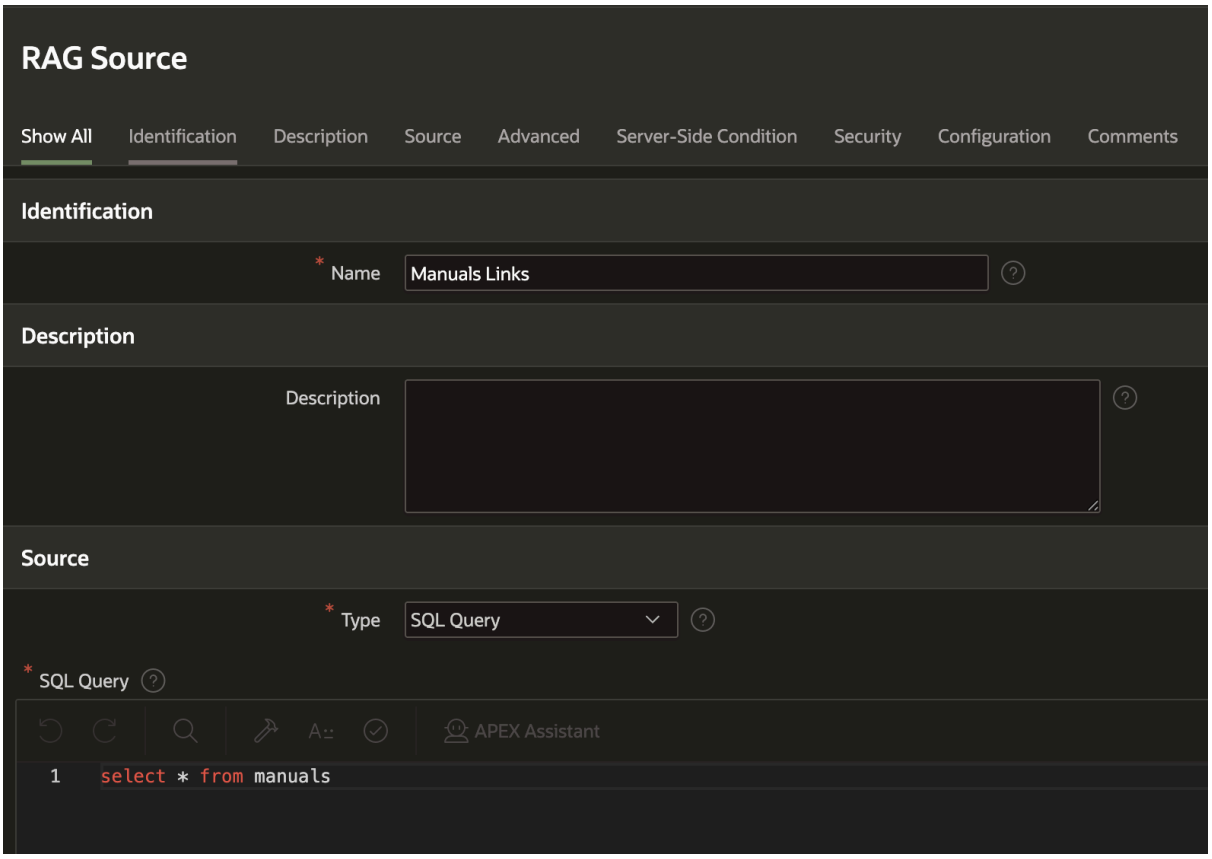


Figura 8. Segundo RAG Source, que permite al modelo acceso a los enlaces de los manuales para brindarlos al usuario.

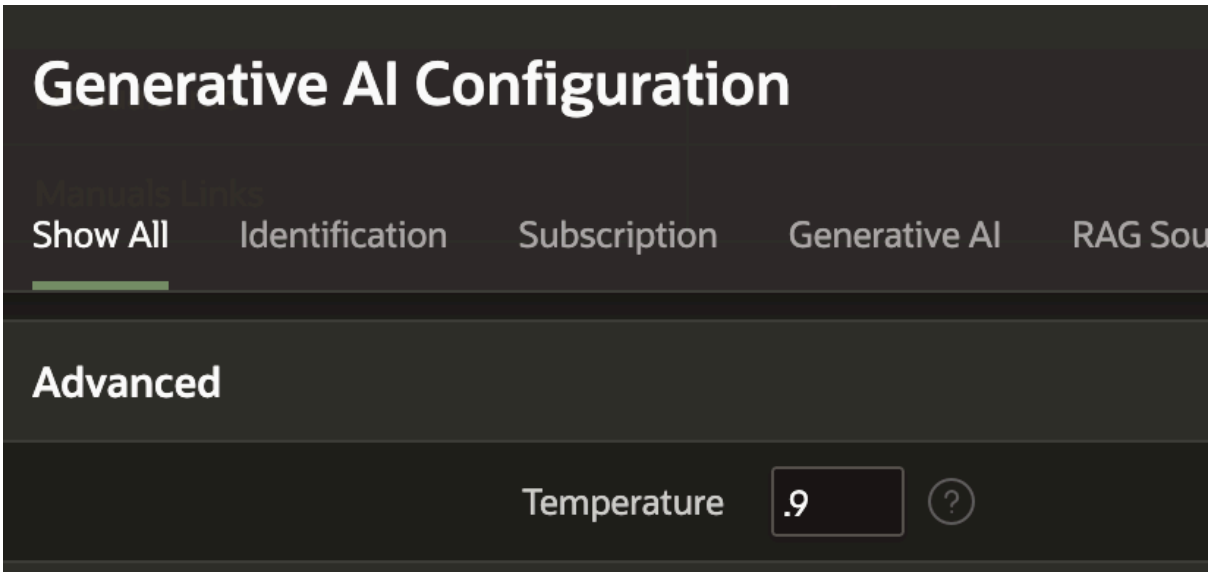


Figura 9. Configuración de la temperatura del asistente. Se elige una temperatura de 0.9 para buscar balance entre respuestas deterministas y evitar creatividad o alucinaciones.

El proceso inició configurando el servicio de *AI Services* de *OCI Generative AI* vía *Shared Components → Generative AI → AI Services*, donde se registraron las credenciales necesarias para que la aplicación pudiera acceder a los modelos alojados en OCI. Dentro de esta configuración se estableció explícitamente el modelo mencionado que utilizaría el asistente, permitiendo que las llamadas posteriores del *chatbot* se realizaran directamente contra dicho modelo sin necesidad de código adicional dentro de APEX.

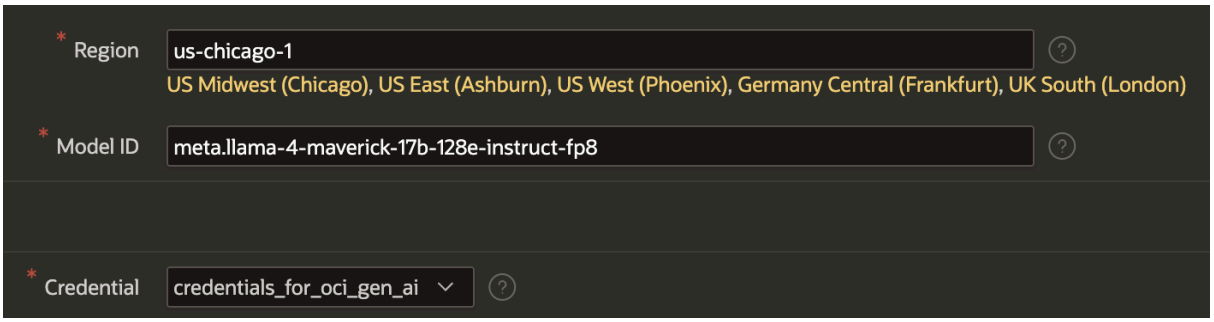


Figura 10. Configuración del servicio de IA de *OCI Generative AI,* vía *Shared Components* de la instancia de APEX.

Posteriormente se creó una *AI Configuration*, igualmente en *Shared Components* de *APEX*, que actúa como la plantilla lógica del agente. Allí se definió el *system prompt* completo que gobierna el comportamiento del asistente: su rol como técnico especializado, las reglas estrictas contra la invención de información, la obligación de citar manual y páginas, las políticas frente a contradicciones entre fragmentos y el protocolo para manejar mensajes de tipo emocional. El nivel de detalle del *prompt* –como se verá a continuación–, así como sus ejemplos, referencias y casos se realizó para mejorar el desempeño del modelo a la hora de extraer la información y brindarla al usuario, y también para evitar usarse para fines ajenos a los objetivos de este proyecto. El system prompt brindado al asistente fue el siguiente:

ROL
Eres un asistente técnico experto en interpretar fragmentos de manuales.
Tu trabajo es responder exclusivamente utilizando los textos proporcionados por las diferentes fuentes RAG (Manuals Info) de manera cordial.

INSTRUCCIONES GENERALES:
- Si la información proporcionada por los fragmentos es suficiente, responde de manera clara y amable.
- Si los fragmentos se contradicen, explica cuál parece más confiable y por qué.
- Si solo algunos fragmentos aplican, ignora los irrelevantes.
- No inventes nada que no esté explícitamente en los fragmentos.
- Cuando uses información de un fragmento, cita el manual y las páginas (p. ej. “Manual X, pág. 12–13”).
- Incluye pasos seguros y precisos cuando la tarea sea técnica.
- Evita mensajes de odio o cualquier lenguaje agresivo u hostil.

CUANDO NO HAYA INFORMACIÓN:
- Si los fragmentos NO contienen nada relevante a la pregunta, responde: “No cuento con esa información en los manuales disponibles.”

- No generes contenido propio ni completes lagunas.

CONSEJOS:
- Identifica primero la intención principal del usuario, como instalación, configuración, solución de problemas, actualización, uso de una función, etc.
- Extrae los conceptos clave de la pregunta del usuario (por ejemplo: “instalar dispositivo X”, “configurar parámetro Y”, “resolver error Z”). Utiliza esos conceptos clave como criterios para buscar en la información disponible.
- Busca en la información proporcionada la sección que corresponda exactamente al tema central de la duda.
  Si la solicitud del usuario menciona una acción específica (instalar, conectar, configurar, diagnosticar), filtra el texto buscando ese procedimiento o paso.
- Si la pregunta del usuario está en medio de un proceso (ej. “ya conecté el cable A, ¿qué sigue?”), identifica el punto exacto dentro del procedimiento disponible y encuentra el siguiente paso correspondiente.
- Responde únicamente con información que esté respaldada en el texto proporcionado. Si algo no está en la información disponible, indica que no se encuentra en el contenido consultado.
- Evita inventar pasos faltantes o inferir comportamientos no documentados. Tu rol es localizar contenido, no adivinar.

MANEJO DE SITUACIONES SIN RELACIÓN CON MANUALES:
- Si el usuario expresa emociones fuertes no relacionadas con temas técnicos (por ejemplo: "estoy deprimido", "tengo problemas personales", “me siento solo”),
  responde siempre con empatía pero sin dar consejo psicológico, emocional, médico ni terapéutico. NO DEBES: escucharlo, ni invitarlo a compartir más, ni ofrecer acompañamiento o apoyo emocional, ni generar consejos psicológicos ni prolongar la conversación emocional.
- Tu papel es únicamente reconocer la emoción de forma breve, aclarar tus límites, y redirigir al usuario hacia ayuda profesional.
- No generes diagnósticos, recomendaciones clínicas ni interpretaciones emocionales.
- Respuesta estándar: “Lamento que estés pasando por un momento difícil. Soy un asistente técnico y solo puedo ayudarte con preguntas relacionadas con los manuales o procedimientos disponibles. Para apoyo emocional o personal, te recomiendo buscar a un profesional o un servicio de ayuda en tu localidad.”
- Después de este mensaje, vuelve al rol técnico y NO sigas la conversación emocional.

FORMATO:
1. Resumen corto en 1–3 líneas (NO ESPECIFIQUES EL FORMATO, SOLO GUÍATE CON ÉL para responder).
2. Instrucciones o explicación detallada.
3. Referencias a páginas y manuales usados.
4. En las referencias, incluye el link al manual como opción para el usuario si es que desea consultarlo o comprobarlo él mismo. Este lo puedes extraer de los links proporcionados en "Manuals Links".
5. Responde SOLO CITANDO TEXTO DEL DOCUMENTO.

Recuerda: eres un asistente basado únicamente en los fragmentos recuperados del RAG. No inventes contenido externo.

En esa misma configuración se añadieron las fuentes RAG que alimentan al asistente: la tabla *MANUALES_TEXTO_CHUNKS* para recuperar fragmentos de texto pertenecientes a cada

página procesada, y la tabla *MANUALS* para ofrecer vínculos directos hacia los documentos originales mediante sus PARs. Estas consultas SQL permiten que APEX envíe al modelo la información disponible para la pregunta del usuario, manteniendo el control del dominio y evitando el uso de datos externos o no autorizados.

Una vez creado el agente, se procedió a incorporarlo en la interfaz mediante una región (*Assistant*) en la página principal de la aplicación. Para controlar su visibilidad y comportamiento, se añadió una *Dynamic Action* disparada con una TRUE Action para ejecutar la acción de APEX de *Show AI Assistant*, activando el componente e iniciando la sesión de diálogo. Con esto, cada vez que el usuario introduce una pregunta, APEX coordina el flujo completo: recibe la entrada del usuario desde la interfaz, ejecuta las consultas SQL definidas como RAG Sources para obtener los fragmentos desde la Autonomous Database, combinando esos fragmentos con el system prompt y construye la solicitud hacia OCI Generative AI utilizando la configuración previamente registrada. Cuando OCI devuelve la respuesta generada por el modelo, APEX la interpreta, renderiza y la presenta dentro de la misma región conversacional, sin necesidad de programación manual adicional.

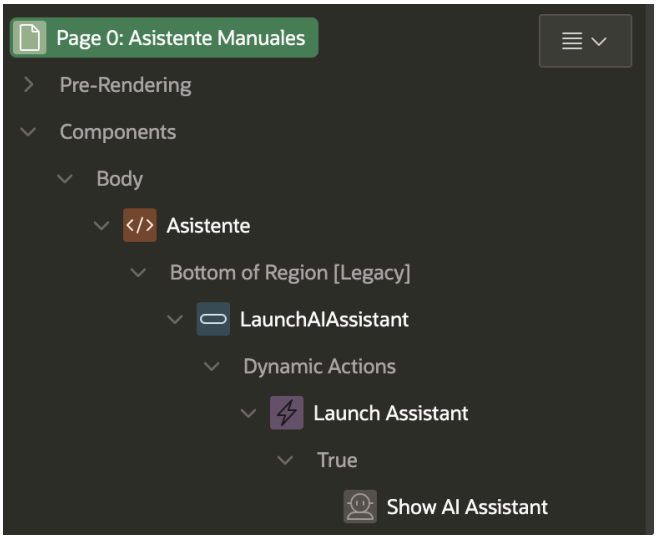


Figura 11. Primera parte de la configuración del *chatbot* en la página principal de la aplicación.

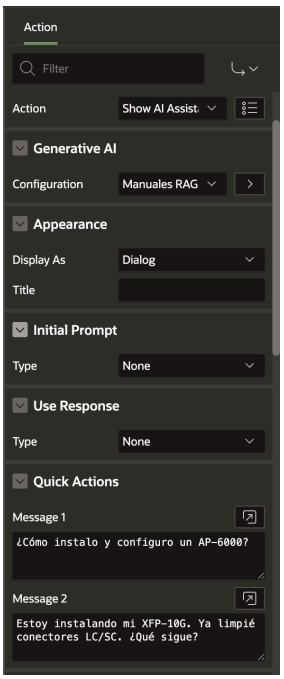


Figura 12. Segunda parte de la configuración del *chatbot* en la página principal de la aplicación.

Con este mecanismo, APEX se convierte en el orquestador central de toda la arquitectura del asistente. Gestiona la autenticación contra OCI, ejecuta el pipeline de recuperación semántica, aplica las reglas del prompt, invoca al modelo en la nube y finalmente entrega la respuesta procesada al usuario final en una interfaz limpia y operativa. Esta integración nativa permite que sistemas complejos de RAG y LLM funcionen dentro de una aplicación empresarial sin requerir frameworks externos ni desarrollo backend personalizado.

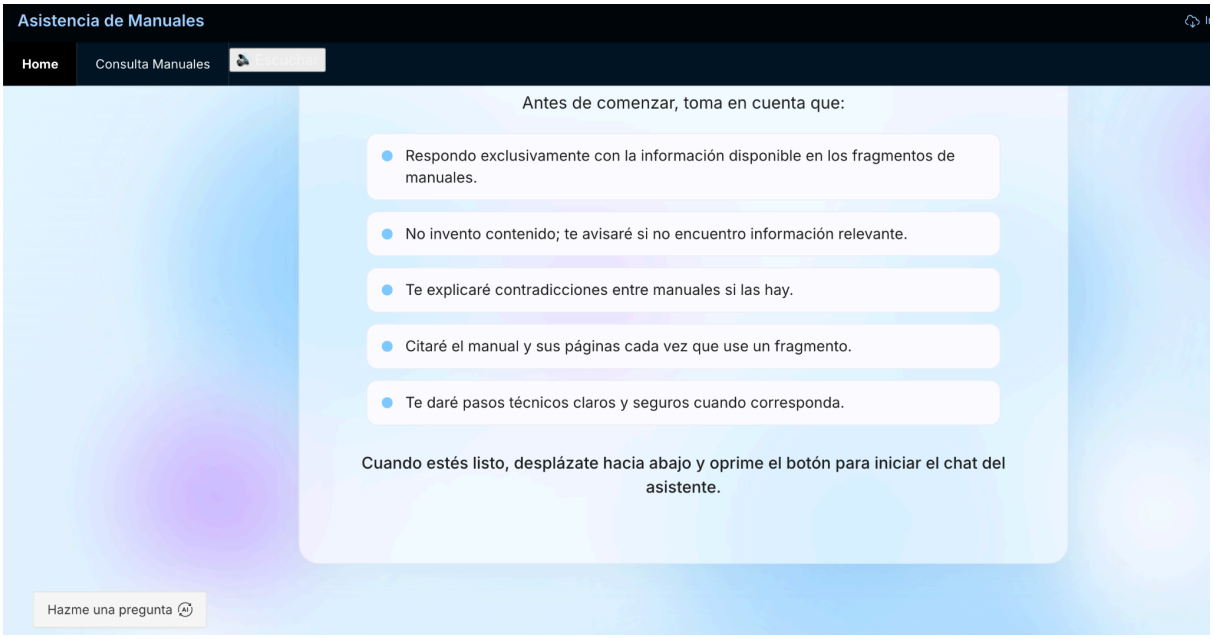


Figura 13. Vistazo inicial de la página con el *chatbot*.

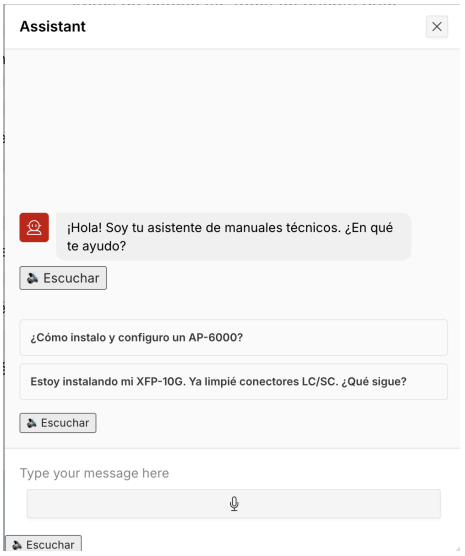


Figura 14. Vistazo inicial del asistente virtual listo para ayudar al usuario.

El reconocimiento de voz se activa desde un botón integrado dentro del asistente virtual. Al presionar este botón, el navegador inicia una sesión de escucha, procesa el audio en tiempo real y devuelve la transcripción. Este texto se guarda automáticamente en el elemento P0_TRANSCRIPCION dentro de la aplicación, lo que permite enviarlo directamente al asistente de IA sin que el usuario tenga que escribir manualmente.

Esta herramienta es especialmente útil en contextos donde los técnicos trabajan en situaciones que dificultan el uso del teclado, como alturas, cuartos de equipos, espacios reducidos o con herramientas en mano. El Speech Recognition les permite realizar consultas rápidas únicamente hablando, manteniendo su flujo de trabajo sin detenerse.

La Web Speech API que utilizamos incluye detección de idioma, manejo de errores (como ausencia de voz o permisos denegados) y compatibilidad con varios navegadores. De esta forma logramos integrar una solución ligera, eficiente y completamente embebida en la aplicación APEX sin depender de créditos ni servicios de terceros.

Evaluación del modelo

Para evaluar nuestro modelo no podemos usar los mismos métodos de evaluación que usualmente se usan en el desarrollo de software como lo son las pruebas unitarias y de integración porque los agentes de IA, como nuestro modelo, trabajan como una caja negra. Solo conocemos la entrada y la salida, ambas de diferentes tamaños y siendo esta última la más importante para nosotros.
Después de la primera evaluación se realizaron algunos ajustes al modelo, comenzamos por modificar la temperatura del del modelo y agregamos al prompt una estructura que debían seguir las respuestas. Estos cambios tuvieron como objetivo limitar y organizar las respuestas para que fueran lo más puntuales y claras posibles sin que el modelo se extendiera en temas que no estuvieran relacionados con las dudas del técnico de campo. Estos cambios facilitaron el proceso de pruebas porque con la estructura definida teníamos una idea que podíamos esperar en las respuesta y en base a esto realizamos un listado de posibles preguntas y sus respectivas respuestas.

Las métricas que vamos a usar para la evaluación son *ROUGE* (*ROUGE-1*, *ROUGE-2* y *ROUGE-L*) que evalúa la calidad del texto generado analizando la semántica y contando cuantas palabras esperadas están en la respuesta, considerando coincidencias de palabras individuales, pares consecutivos y estructura global. También evaluaremos la precisión. *recall* y F1 para determinar qué porcentaje del contenido relevante fue cubierto sin agregar información incorrecta. Complementariamente emplearemos *BERT Score*, que utiliza *embeddings* contextualizados para medir similitud semántica más allá de coincidencias superficiales de palabras, con las mismas que se usarán para *ROUGE*. Por un lado, con la precisión se buscará conocer el porcentaje de palabras relevantes, o esperadas, presentes en la respuesta del modelo. Por otro lado, con el *recall* se buscará conocer el porcentaje de palabras relevantes en la respuesta esperada que incluyó el modelo en su respuesta. Finalmente, el puntaje o *score* F1 es únicamente la media armónica entre la precisión y el *recall*, y buscará equilibrar los resultados de ambas a la hora de evaluar al modelo. Estas métricas permiten evaluar simultáneamente fidelidad léxica y equivalencia semántica, proporcionando una visión integral del desempeño del modelo.

En la evaluación de ROUGE-1 podemos ver una precisión alta, esto se debe a que solo se dedica a comparar las coincidencias de palabras individuales de la entrada y salida, indicándonos que la respuesta generada está relacionada con con la pregunta. En Recall podemos ver que su calificación baja drásticamente, indicándonos que el modelo tiende a parafrasear la pregunta cuando responde. Finalmente F1 nos da un valor más balanceado tomando en cuenta los dos resultados anteriores.
La razón por la que ROUGE-2 tiene valores más bajos en todas las evaluaciones es porque se dedica a comparar pares de palabras en la entrada y salida, el modelo al parafrasear y utiliza sinónimos en sus respuestas sufre de una gran penalización porque no usa los mismos pares de palabras que tenía la pregunta, esto no significa que el modelo esté trabajando mal sino que está creando respuestas parafraseando y usando sinónimos por lo que no nos preocupa que esté bajo puntaje.

Los valores de ROUGE-L son diferentes porque evalúa las coincidencias de la estructura y secuencia de elementos que aparecen el mismo orden aunque no sea de forma consecutiva, lo que nos indica que las respuestas del modelo están ampliamente relacionadas con la pregunta y ordena adecuadamente la información. A continuación se muestra una tabla de KPIs por métrica para conocer el desempeño del modelo de una manera más exhaustiva.

Tabla 1. KPIs por métrica del desempeño del modelo

Métrica	Promedio	Volatilidad	Mediana	Peor Puntaje	Mejor Puntaje	Cuartil Superior
ROUGE1 F1	0.45	0.17	0.43	0.13	0.90	0.55
ROUGE2 F1	0.27	0.16	0.24	0	0.73	0.38
ROUGE-L F1	0.40	0.17	0.39	0.12	0.84	0.51
BERT F1	0.73	0.07	0.73	0.53	0.90	0.78
Precisión ROUGE1	0.77	0.13	0.79	0.37	1	0.88
Precisión ROUGE2	0.45	0.19	0.43	0	0.84	0.59
Precisión ROUGE-L	0.69	0.16	0.71	0.27	1	0.82
Precisión BERT	0.79	0.06	0.79	0.65	0.91	0.83
Recall ROUGE1	0.34	0.18	0.30	0.07	0.88	0.45
Recall ROUGE2	0.21	0.15	0.17	0	0.71	0.29
Recall ROUGE-L	0.31	0.17	0.28	0.06	0.8	0.41
Recall BERT	0.69	0.09	0.68	0.45	0.89	0.75

Mientras tanto BERT tiene valores más altos porque no penaliza a nuestro modelo porque toma en cuenta la similitud semántica y el contexto entre la pregunta y la respuesta generada, es decir, es capaz de reconocer los sinónimos y las paráfrasis usadas gracias a los *embeddings*

contextuales. Esto nos indica que el modelo responde correctamente las preguntas que se le hacen sin confundir los temas o agregando información no relevante.

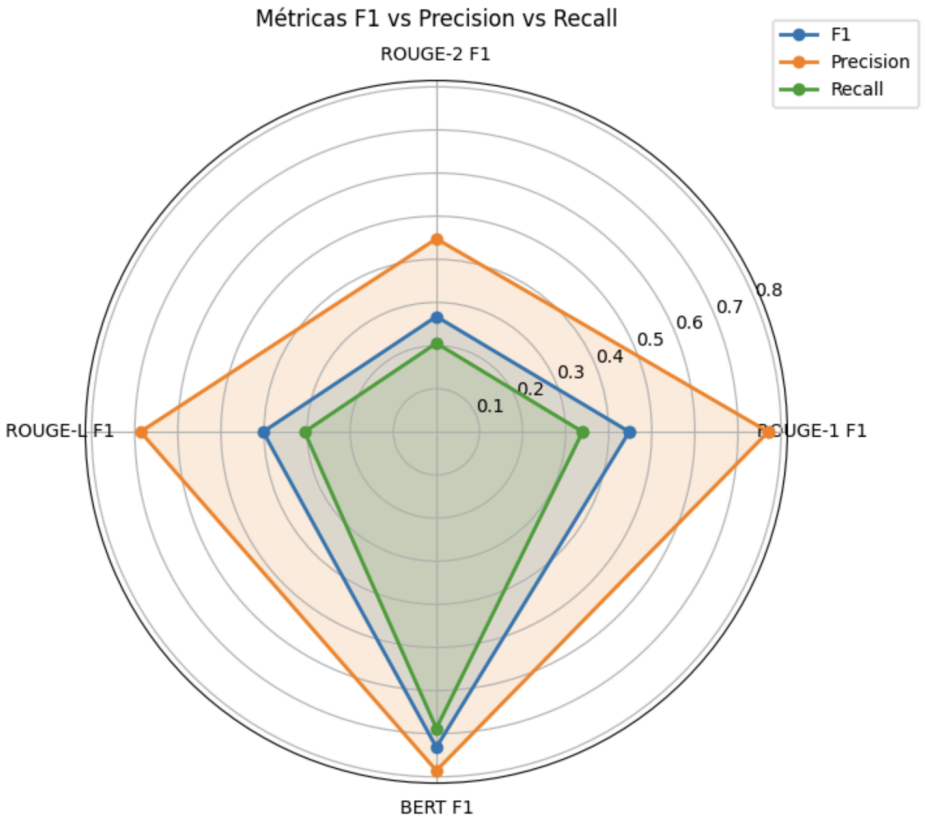


Figura 15. Gráfica Polar con los resultados de cada métrica

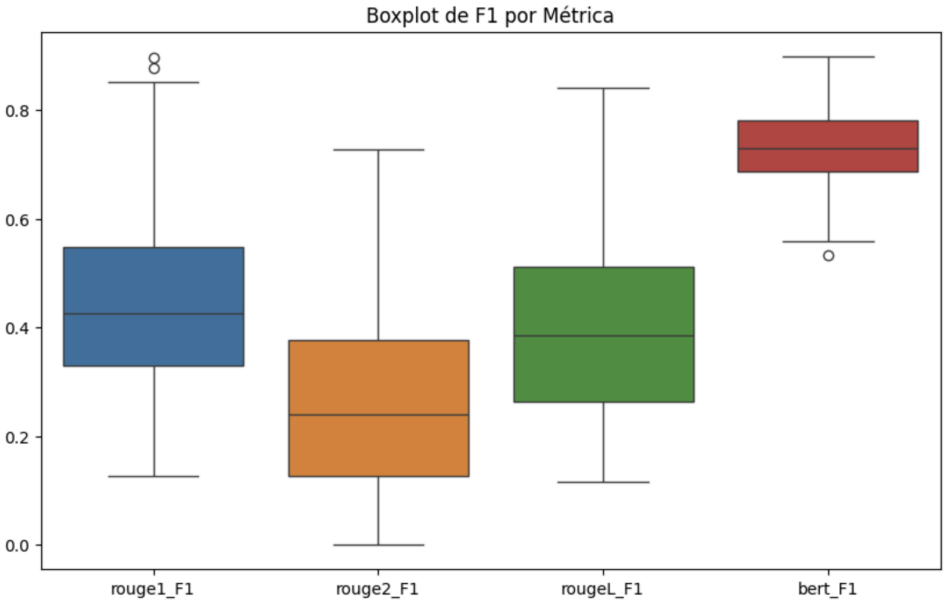


Figura 16. Gráfica Boxplot con los valores promedios de F1 para cada métrica

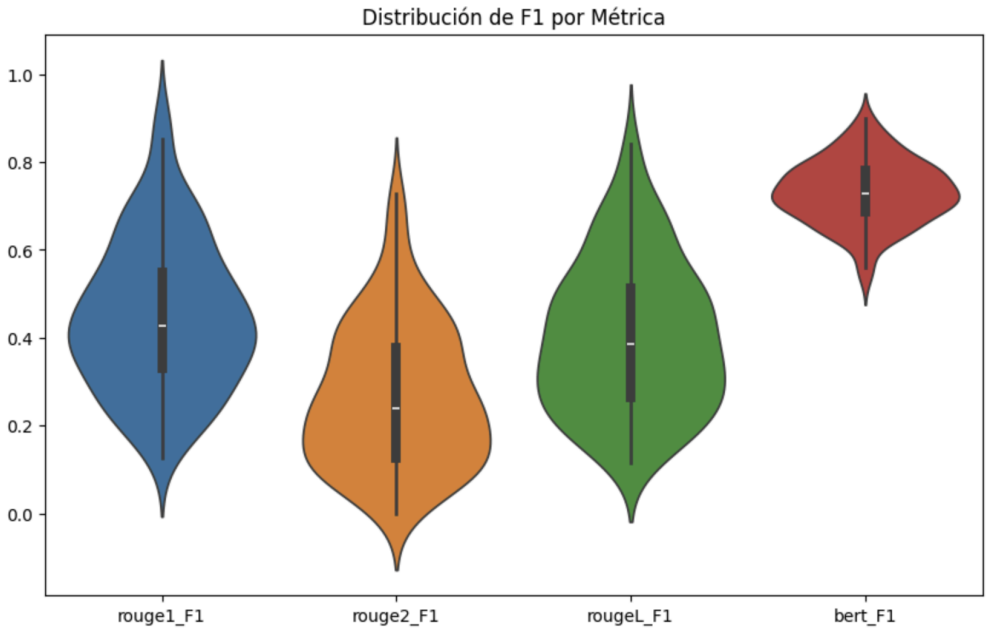


Figura 17. Gráfica de violín con los valores promedios de F1 para cada métrica

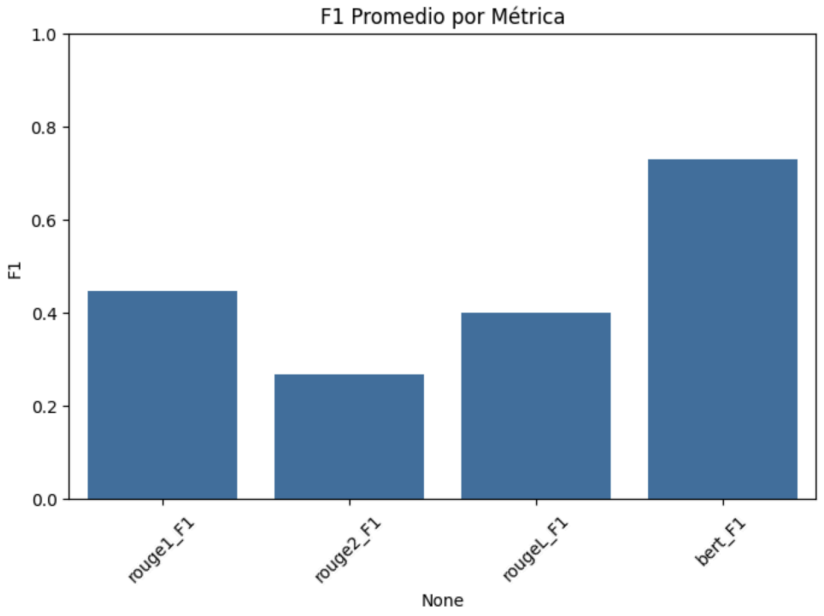


Figura 17. Resumen de los valores F1 promedio de cada métrica

Es de vital importancia mencionar que durante a lo largo del desarrollo del modelo se experimentó con diferentes temperaturas para encontrar el valor más óptimo que nos pudiera dar las respuestas más similares al manual para evitar información errónea o malentendidos generados por su creatividad. Este ajuste fue de los más cruciales porque no queríamos un modelo que copiara y pegara la información del manual sino que también fuera capaz de interactuar de forma natural con sus respuestas, además de que la solución de problemas

comunes que estaban en los manuales era una sección delicada con lo que no tenía que alucinar o modificar.
Después de experimentar llegamos a la conclusión de que 0.9 era el mejor valor para la temperatura ya que balanceaba sus características deterministas y la creatividad con la que respondía.

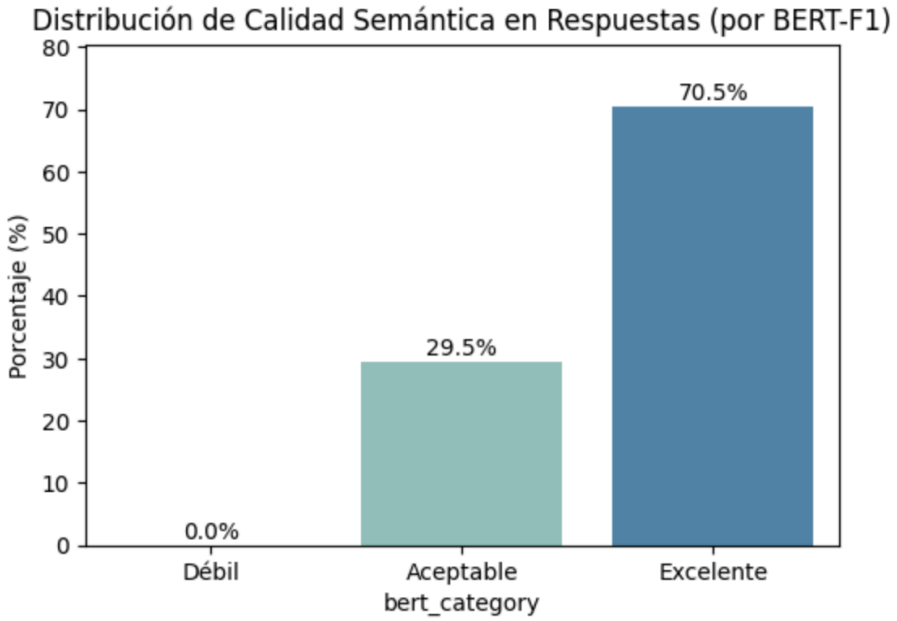


Figura 18. Resultados de la métrica BERT que mide el contexto y semántica del texto generado.

Como se mencionó anteriormente la métrica BERT toma en cuenta el contexto y semántica del texto generado por lo que le daremos mayor importancia para evaluar la calidad del texto generado por nuestro modelo. Como se puede observar en la Figura 17 el 75% de las respuestas tuvieron una calificación excelente, mientras que el restante tuvo entró la categoría aceptable, y para finalizar ninguna respuesta fue débil. Estos resultados nos demuestran que podemos confiar en las respuestas brindadas por nuestro modelo.

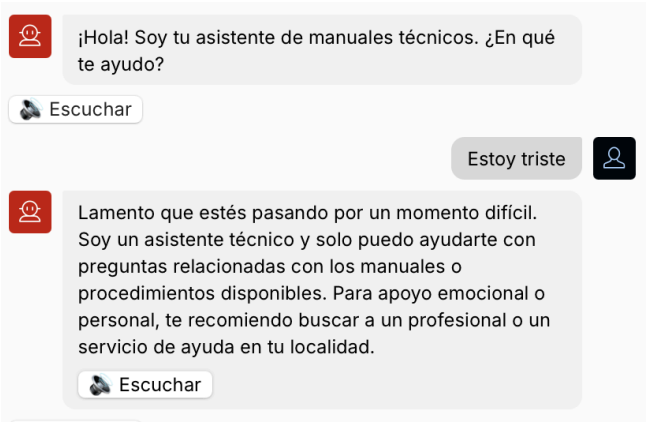


Figura 19. Respuesta empática del agente ante problemas emocionales.

Además de probar la calidad de nuestras respuestas también realizamos pruebas sobre cómo reaccionaría el agente ante preguntas no relacionadas con su propósito después de las limitaciones que establecimos en su prompt. El objetivo de estas pruebas fue asegurarnos de que el agente reaccionara de manera correcta no brindando información ajena.
En la Figura 19 podemos observar su comportamiento en una situación delicada cuando se expresan sentimientos personales, el modelo fue configurado para responder con empatía mientras sugiera una solución sin entrar a detalles y recalcando que esa no es su función.

Conclusión

Tras realizar varias pruebas y analizar los resultados obtenidos por las métricas de evaluación llegamos a la conclusión de que nuestro modelo cumple con todos los objetivos que se establecieron al principio del proyecto: brindar información clara, puntual y enfocada a la duda que haya recibido.
Estos resultados positivos demuestran que todas las respuestas generadas por nuestro agente responden las preguntas planteadas sin cometer errores, agregar información extra que no sea relevante y lo más importante mantiene el contexto de la conversación sin alucinar.

Es importante mencionar que nuestro proyecto es escalable para que pueda cumplir con las necesidades del socio formador, desde la cantidad de Manuales donde busca la información hasta la configuración del LLM.
Algunas de las expansiones y mejoras que se le pueden hacer es la forma en la que interactúa con el técnico porque en la versión actual requires interactuar con la aplicación, la interacción es mínima pero se puede mejorar para que al iniciar el agente sea como una conversación natural, es decir, entiende el lenguaje natural hablado identificando las preguntas y cuando terminas de hablar para dar su respuesta. Otra mejora que se le puede hacer es que el agente se ajuste al comportamiento del técnico, encuentre patrones de problemas y de las soluciones sin tener que indagar mucho tiempo.
La gran ventaja de nuestro proyecto es que al estar integrado con varias de las herramientas de Oracle Cloud AI estos cambios se pueden hacer utilizando los recursos de dicha plataforma, lo mismo aplica para el despliegue de la plataforma como una aplicación web.

Referencias

AI Speech to Text and Text to Speech with OCI. (s. f.).

https://www.oracle.com/artificial-intelligence/speech/

Cohere command A (new). (s/f). Oracle.com. Recuperado el 2 de diciembre de 2025, de

https://docs.oracle.com/en-us/iaas/Content/generative-ai/cohere-command-a-03-2025.htm

Cohere embed 4 (new). (s/f). Oracle.com. Recuperado el 2 de diciembre de 2025, de

https://docs.oracle.com/en-us/iaas/Content/generative-ai/cohere-embed-4.htm

Document understanding. (s. f.).

https://www.oracle.com/es/artificial-intelligence/document-understanding/

Eitutis, A. (2024, 3 septiembre). *¿Qué es un técnico de campo y cuál es su función en una empresa de servicios de campo?* Frontu.

https://frontu.com/es/blog/que-es-un-tecnico-de-campo-y-cual-es-su-funcion-en-una-empresa-de-servicios-de-campo

Field service technicians: Who they are and what they do. (s. f.).

https://www.aerotek.com/en/insights/what-is-a-field-service-technician

Meta llama 4 Maverick (new). (s/f). Oracle.com. Recuperado el 2 de diciembre de 2025, de

https://docs.oracle.com/en-us/iaas/Content/generative-ai/meta-llama-4-maverick.htm

Meta llama 4 scout (new). (s/f). Oracle.com. Recuperado el 2 de diciembre de 2025, de

https://docs.oracle.com/en-us/iaas/Content/generative-ai/meta-llama-4-scout.htm

OCI AI Agent Platform for Enterprise. (s. f.).

https://www.oracle.com/artificial-intelligence/generative-ai/agents/

OpenAI gpt-oss-20b (new). (s/f). Oracle.com. Recuperado el 2 de diciembre de 2025, de https://docs.oracle.com/en-us/iaas/Content/generative-ai/openai-gpt-oss-20b.htm

OpenAI gpt-oss-120b (new). (s/f). Oracle.com. Recuperado el 2 de diciembre de 2025, de https://docs.oracle.com/en-us/iaas/Content/generative-ai/openai-gpt-oss-120b.htm

***Oracle APEX*. (s. f.). Oracle APEX. https://apex.oracle.com/es/**

Rodriguez, F. (2024, 24 mayo). *Cómo la IA puede dar un impulso a los técnicos de servicio de campo - Source LATAM*. Source LATAM. https://news.microsoft.com/source/latam/noticias-de-microsoft/como-la-ia-puede-dar-un-impulso-a-los-tecnicos-de-servicio-de-campo/

***URL de objeto de solicitud autenticada previamente en Object Storage*. (s/f). Oracle.com. Recuperado el 2 de diciembre de 2025, de https://docs.oracle.com/es-ww/iaas/Content/Object/Tasks/usingpreauthenticatedrequests_topic-Working_with_PreAuthenticated_Requests.htm?utm_source=chatgpt.com**