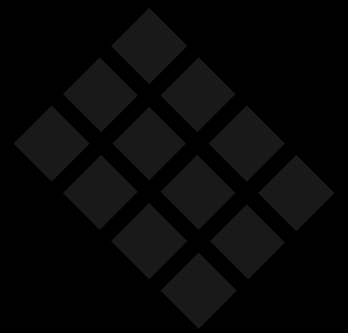




```
POST https://llm.tirescue.com/v1/chat/completions

// Tu aplicación, con la misma librería que usa para OpenAI
{
  "model": "rescue-main",
  "messages": [
    { "role": "system", "content": "Eres el asistente de ventas." },
    { "role": "user", "content": "¿Cuántos pedidos quedaron pendientes de facturar esta semana?" }
  ],
  "tools": [ { "name": "consultar_pedidos", ... } ]
}

// Respuesta · 0,9 s · modelo privado de T.I. Rescue
{ "tool_calls": [ { "name": "consultar_pedidos",
  "arguments": { "estado": "pendiente", "semana": "actual" } } ] }

// ...y con el resultado de tu sistema:
"Quedan 14 pedidos pendientes por $38,2 millones.
Tres superan los 5 días: #2249, #2484 y #2615."
```



# FICHA TÉCNICA RESCUE AI

Inteligencia artificial privada para tus aplicaciones.



**T.I. RESCUE**  
SEGURIDAD INFORMÁTICA

## ISO 9001

Seguridad informática premium, respaldada por el reconocimiento internacional ISO 9001.



SC-CER900888

## ISO 27001

Defendiendo tus datos con rigor internacional: certificación ISO/IEC 27001.



SI-CER900889

## ISO 20000-1

Excelencia en gestión de servicios de TI, fundamentada en estándares mundiales de calidad.



TI-CER988544

## ISO 22301

Continuidad del negocio sin interrupciones, conforme a estándares internacionales.



CO-CNCER988779

# CUATRO CERTIFICACIONES ISO DEFINEN NUESTRA SUPERIORIDAD

Nuestro compromiso con la ciberseguridad trasciende la simple retórica. Es tangible y está validado por cuatro prestigiosas certificaciones **ISO 9001, 27001, 20000-1 y 22301**. No son solo siglas: son el testimonio de nuestra dedicación a ofrecer servicios de primera clase y representan nuestra promesa de calidad, seguridad, **eficiencia y resiliencia**. Al elegirnos no solo escoges un proveedor de ciberseguridad, sino un aliado certificado que entiende y cumple los más altos estándares internacionales.

## COMPROMISO INTEGRAL: MÁS ALLÁ DE LA SEGURIDAD BÁSICA

Con nuestras certificaciones garantizamos **calidad, seguridad de la información, excelencia en TI y continuidad del negocio**. Representan nuestro compromiso, profesionalismo e integridad. Elegirnos es optar por una ciberseguridad de élite y con respaldo.



SC-CER900888

SI-CER900889

TI-CER988544

CO-CNCER988779

# SOMOS UN ALIADO ESTRATÉGICO

Para la seguridad informática e  
infraestructura tecnológica de las  
empresas.

## EMPRESA FUNDADA EN 2015 CON BASE EN COLOMBIA

Con operación en **todo el mundo**

### MISIÓN

Nos especializamos en el soporte y la seguridad informática con los más altos estándares internacionales, incluidas las normas **ISO 9001, 27001, 22301 y 20000-1**. Nuestra misión es garantizar la continuidad operativa de empresas de todos los tamaños con soluciones tecnológicas eficientes y seguras, y servicios personalizados que se ajustan a las necesidades de cada cliente, **evitando interrupciones y promoviendo su éxito.**

### VISIÓN

Ser líderes globales en administración y seguridad informática, con soluciones personalizadas y eficientes que garanticen la continuidad operativa. Estamos comprometidos con **la innovación y el cumplimiento de los estándares de calidad y seguridad más rigurosos**, para ser siempre la primera opción de nuestros clientes.

**Contáctanos**  
**info@tirescue.com.co**  
**www.tirescue.com**

# TU SEGURIDAD, NUESTRA PRIORIDAD

## Soluciones adaptadas

A la medida de cada cliente, sin importar si la empresa es pequeña o grande.

## Actualización constante

Soluciones inmediatas para prevenir pérdidas de información e ir siempre un paso adelante.

## Comunicación directa

Sin filas ni turnos.

## Ahorro en hardware

Lo que se implementa en software.

## Soporte 24/7

Acompañamiento permanente cuando lo necesitas.

## Personal atento

Un equipo certificado que conoce tu operación.

## NUESTROS VALIOSOS CLIENTES

Asociaciones basadas en la confianza y la excelencia



SILK DUBAI



Grupo OET

Alevilla



ipanū

Oral-Plus  
SALUD Y BELLEZA EN TU SONRISA



anquímico

Ame

juju  
Soluciones amigables

Clínica mery álvarez  
-Desde 1999-



artmedica

PKF  
Integral

LLERAS PARK  
CONCEPT HOTEL

AUDITORÍAS  
E INFORMES

Ame

Emprestur  
Más que Transporte

hra

Grupo Agrosiete  
S.A.S.

COOPEVIAN  
SEGURIDAD PRIVADA

HA  
Bicicletas



SP  
Soluciones & Partes  
S.A.S.

TELSAT

Sary

# IA POR API, EN SERVIDORES PRIVADOS

PLATAFORMA RESCUE AI

## CONCEPTO FUNDAMENTAL UN MODELO DE IA QUE TRABAJA DENTRO DE TUS SISTEMAS

**Rescue AI** pone un modelo de lenguaje de última generación al servicio de tus aplicaciones: tu ERP, tu CRM, tu mesa de ayuda o tu sitio web lo consultan por API y el modelo responde, consulta tus datos a través de las funciones que tú defines y entrega el resultado en segundos.

Todo corre en infraestructura administrada por T.I. Rescue. No hay cobro por cada palabra ni facturas sorpresa: contratas una capacidad fija, como un servidor, y la amplías cuando tu uso crece.

### Compatible con OpenAI

La misma API y las mismas librerías. Cambias la URL y la clave; tu código sigue igual.

### Datos **privados**

Tus consultas no entrenan modelos ni se comparten con terceros.

### Uso **ilimitado**

Sin cobro por token. Contratas capacidad, como un servidor.

### Capacidad **garantizada**

Consultas en paralelo reservadas para tu empresa.

### Herramientas y **visión**

El modelo llama a tus funciones, lee imágenes y responde en JSON.

### Escala en **minutos**

Subes de plan sin cambiar nada en tu aplicación.

# CÓMO FUNCIONA DE LA PREGUNTA A LA RESPUESTA

Tu aplicación envía la consulta con su clave. La pasarela de Rescue AI verifica la clave y la capacidad contratada, elige el modelo indicado y devuelve la respuesta por streaming. Si el modelo necesita un dato de tu sistema, lo pide con una llamada a función y tú se lo entregas: él nunca toca tu base de datos directamente.



**131 K**

TOKENS DE CONTEXTO  
POR CONSULTA

**96 GB**

DE MEMORIA GPU POR  
MODELO

**5**

CAPACIDADES: TEXTO,  
RAZONAMIENTO, VISIÓN,  
IMÁGENES, BÚSQUEDA

**0**

COBRO POR TOKEN

- Respuestas por **streaming**
- Salida en **JSON** estructurado
- Panel de uso por **clave**
- Llamadas a **funciones** (tools)
- Modelo de **respaldo** automático
- Varias claves por **empresa**

# PLANES

## CAPACIDAD FIJA, USO ILIMITADO

Funciona como el hosting: contratas una capacidad y la usas sin contar tokens. Cuando necesitas más, subimos el plan el mismo día y se factura la diferencia. Precios mensuales en pesos colombianos.

|                       | Plan S                                                                                          | Plan M                                              | Plan L                                                                |
|-----------------------|-------------------------------------------------------------------------------------------------|-----------------------------------------------------|-----------------------------------------------------------------------|
| Consultas en paralelo | 2<br>conversaciones simultáneas                                                                 | 5<br>conversaciones simultáneas                     | 10<br>conversaciones simultáneas                                      |
| Tokens por minuto     | 300 K                                                                                           | 750 K                                               | 1,5 M                                                                 |
| Consultas por minuto  | 30                                                                                              | 60                                                  | 120                                                                   |
| Tokens al mes         | Ilimitados                                                                                      | Ilimitados                                          | Ilimitados                                                            |
| Uso de referencia     | Hasta 10 usuarios o un asistente de uso ocasional                                               | Hasta 40 usuarios o un asistente conectado a tu ERP | Hasta 80 usuarios o procesos automáticos                              |
| Modelos               | Texto, razonamiento, visión, generación de imágenes y búsqueda semántica (ver página siguiente) |                                                     |                                                                       |
| Claves API            | 1                                                                                               | 3                                                   | 10                                                                    |
| Soporte               | Correo, horario hábil                                                                           | Correo y WhatsApp, horario hábil                    | Correo y WhatsApp, horario hábil, ingeniero asignado · panel en línea |
| Mensualidad           | \$ 900.000<br>COP / mes                                                                         | \$ 1.700.000<br>COP / mes                           | \$ 3.000.000<br>COP / mes                                             |

### Cómo se amplía

Si tus consultas empiezan a esperar turno, lo ves en el panel y nosotros también. Subimos el plan en minutos, **sin cambiar nada en tu aplicación**, y facturamos la diferencia desde ese día.

### Uso razonable

Ilimitado significa personas y aplicaciones trabajando. Procesos masivos desatendidos (por ejemplo, reprocesar toda una base de datos cada noche) se dimensionan aparte, **con una cotización a la medida**.

### ¿Necesitas más capacidad o un entorno exclusivo?

Para volúmenes superiores al Plan L, aislamiento total o un modelo ajustado a tu empresa, preparamos una propuesta a la medida.

# MODELOS Y CAPACIDADES

## UN NOMBRE POR TAREA

Eliges el modelo por lo que necesitas, no por su versión. Cuando actualizamos un modelo, el nombre se mantiene y tu aplicación sigue funcionando. Todos están incluidos en los tres planes.

### rescue-main

TEXTO · 131 K DE CONTEXTO

Modelo principal: redacción, análisis, resúmenes, asistentes conectados a tus datos. Lee documentos largos completos.

### rescue-fast

RÁPIDO · 32 K DE CONTEXTO

Para clasificar, extraer datos y responder en volumen con la menor latencia.

### rescue-reason

RAZONAMIENTO · 131 K

Piensa antes de responder: cálculos, comparaciones y decisiones con varios pasos.

### rescue-vision

VISIÓN · IMÁGENES Y PDF

Describe fotos, lee facturas y pantallas, extrae tablas de documentos escaneados.

### rescue-image

GENERACIÓN DE IMÁGENES

Crea imágenes a partir de texto para piezas, catálogos y contenido.

### rescue-embed

BÚSQUEDA SEMÁNTICA

Convierte textos en vectores para buscar por significado en tus documentos (RAG).

## PARA EL EQUIPO DE DESARROLLO

API OpenAI v1

Streaming

Tool calling

JSON mode

Python · Node · PHP · .NET

LangChain · LlamaIndex

Open WebUI

Cualquier SDK de OpenAI funciona cambiando **base\_url** y **api\_key**. Entregamos ejemplos de integración y acompañamos la primera conexión.

# POR QUÉ CAPACIDAD Y NO TOKENS COMO CONTRATAR UN SERVIDOR

Los proveedores en la nube cobran por cada token que entra y sale. El resultado es una factura que nadie puede prever: un asistente conectado a datos envía decenas de miles de tokens por cada pregunta. Rescue AI usa el modelo del hosting: una capacidad reservada, un precio fijo y ampliación cuando haga falta.

| En un servidor            | En Rescue AI                                                                                               |
|---------------------------|------------------------------------------------------------------------------------------------------------|
| Núcleos de CPU            | <b>Consultas en paralelo:</b> cuántas conversaciones atiende a la vez                                      |
| Memoria RAM               | <b>Tokens por minuto:</b> cuánto contexto puede mover                                                      |
| Tráfico ilimitado         | <b>Tokens al mes ilimitados:</b> no se cuentan ni se facturan                                              |
| El servidor se pone lento | La aplicación recibe <b>«espera un momento»</b> (HTTP 429) y reintenta: la señal de que toca subir de plan |
| Ampliar CPU y RAM         | <b>Subir de plan</b> en minutos, sin reinicios ni cambios de código                                        |
| Panel del hosting         | <b>Panel de uso</b> por clave: consultas, tokens, errores y picos                                          |

## Presupuesto fijo

Finanzas sabe cuánto cuesta cada mes antes de empezar. Los picos de uso no cambian la factura.

## Sin sorpresas técnicas

Los límites se aplican en la pasarela, no en tu código. Nadie en tu empresa puede generar un gasto inesperado.

## Comparable

El Plan M cuesta lo que una suscripción de ChatGPT para 20 personas, pero conectado a tus sistemas y con tus datos en privado.

## Un solo proveedor

Modelo, infraestructura y soporte con el mismo equipo que administra tu seguridad y tu mesa de ayuda.

# FICHA TÉCNICA

## SEGURIDAD Y ESPECIFICACIONES

|                  |                                                                                                                                                                                                                       |
|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Acceso           | Punto único <b>https://llm.tirescue.com/v1</b> con TLS. Clave API por aplicación, revocable al instante, con los modelos permitidos y la capacidad del plan.                                                          |
| Compatibilidad   | API de OpenAI (chat completions, embeddings, imágenes). Funciona con los SDK oficiales y con LangChain, LlamaIndex, n8n y herramientas similares.                                                                     |
| Privacidad       | Las consultas no se usan para entrenar modelos ni se envían a terceros. Modelos abiertos ejecutados en infraestructura administrada por T.I. Rescue.                                                                  |
| Protección       | Límites de concurrencia y de tokens por clave aplicados en la pasarela · guardas de contenido configurables por aplicación · registro de cada consulta (sin almacenar el contenido, salvo que el cliente lo solicite) |
| Disponibilidad   | Dos modelos en servidores distintos con respaldo cruzado automático · monitoreo 24/7 · ventanas de mantenimiento avisadas con 48 h                                                                                    |
| Infraestructura  | GPU NVIDIA RTX PRO 6000 (96 GB) dedicadas a inferencia · vLLM · pasarela LiteLLM · centro de datos europeo con certificación ISO 27001                                                                                |
| Modelos          | Familia Qwen 3 (texto, razonamiento y visión) · Qwen-Image · BGE-M3 (embeddings y reordenamiento). Actualizados sin cambiar los nombres de la API.                                                                    |
| Contexto         | Hasta 131 072 tokens por consulta en rescue-main y rescue-reason · 32 768 en rescue-fast                                                                                                                              |
| Panel y reportes | Uso por clave y por modelo, picos de concurrencia y errores · reporte mensual en PDF · alertas al acercarse al límite del plan                                                                                        |
| Facturación      | Mensual, en pesos colombianos · cambio de plan en cualquier momento, prorrateado · ajuste anual por tasa de cambio                                                                                                    |
| Soporte          | Correo y WhatsApp en horario hábil · acompañamiento en la primera integración · ingeniero asignado en el Plan L                                                                                                       |

# INTELIGENCIA ARTIFICIAL QUE TRABAJA PARA TI



**T.I RESCUE**  
SEGURIDAD INFORMATICA

## CONTÁCTANOS

---

Pide una clave de prueba de Rescue AI y conecta tu primera aplicación esta semana, con nuestro acompañamiento.

---

- Cra. 65 #8B-91 Ofic. 485, Guayabal, Medellín, Antioquia • WhatsApp: +57 315 893 7873
- Teléfono fijo: +57 4 6073303 • [info@tirescue.com.co](mailto:info@tirescue.com.co) • [www.tirescue.com](http://www.tirescue.com)

**WWW.TIRESCUE.COM**  
Expertos en ciberseguridad