VA360

CLASE PRÁCTICA

RAG, Vectores, Embeddings y Chat con tus Datos en n8n

Aprende a crear tu propio asistente de IA que responde preguntas sobre tus documentos reales, sin tecnicismos y paso a paso.

HERRAMIENTAS: n8n Pinecone OpenAI

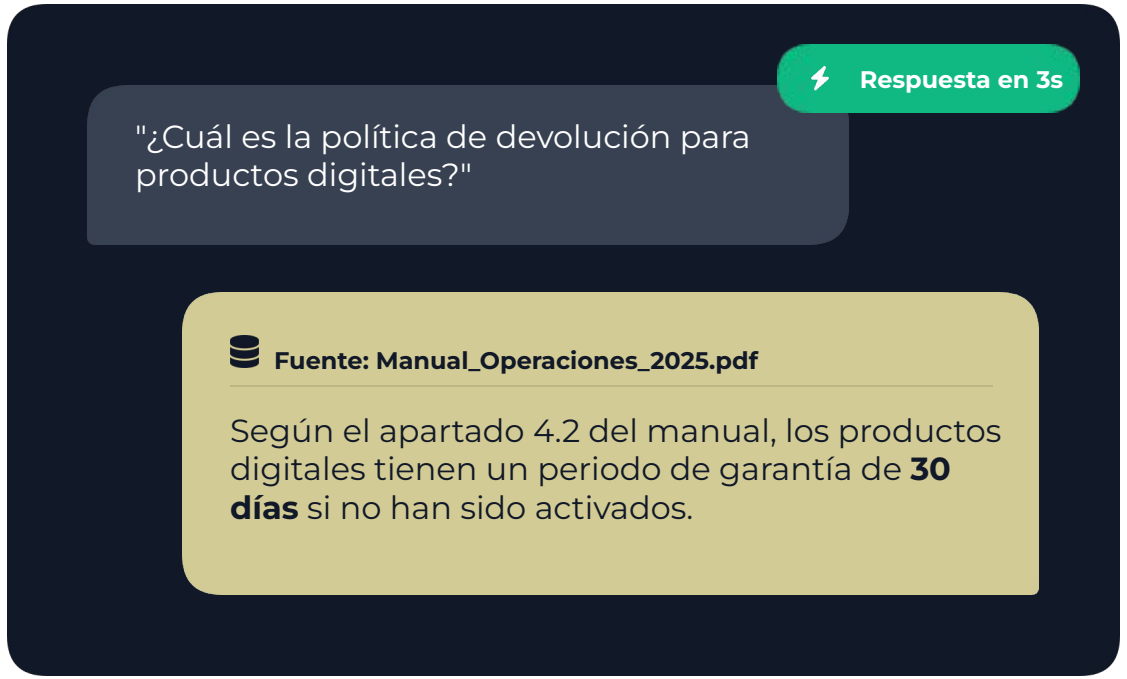
LA GRAN PREGUNTA

¿Y si pudieras hablar con tus documentos?

Imagina que tienes **500 PDFs** de tu empresa: manuales, contratos, FAQs y un cliente te pregunta algo específico.

Método Tradicional:

Buscar manualmente en carpetas, abrir archivos, CTRL+F...
Tiempo estimado: 15-30 minutos

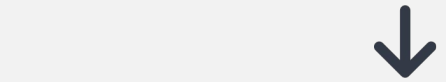


Responde con **TUS DATOS**, no con información de Internet.

PUNTO CLAVE

ChatGPT sabe mucho, pero NO sabe nada de TU empresa.

Hoy voy a enseñarte.



CONCEPTOS: EL TRADUCTOR

¿Qué es un Embedding?

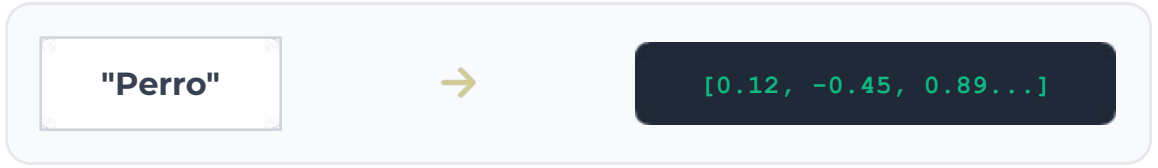
Definición Simple

Los ordenadores no entienden palabras, solo números. Un **Embedding** es un traductor que convierte texto en una lista de números (coordenadas) que representan su **significado**.

Analogía GPS

Tu dirección postal ("Calle Mayor 5") se traduce a coordenadas GPS (36.75, -5.81).
Si vives cerca de alguien, vuestros números GPS serán muy parecidos. Si vives en otro país, serán muy diferentes.

Proceso de Traducción



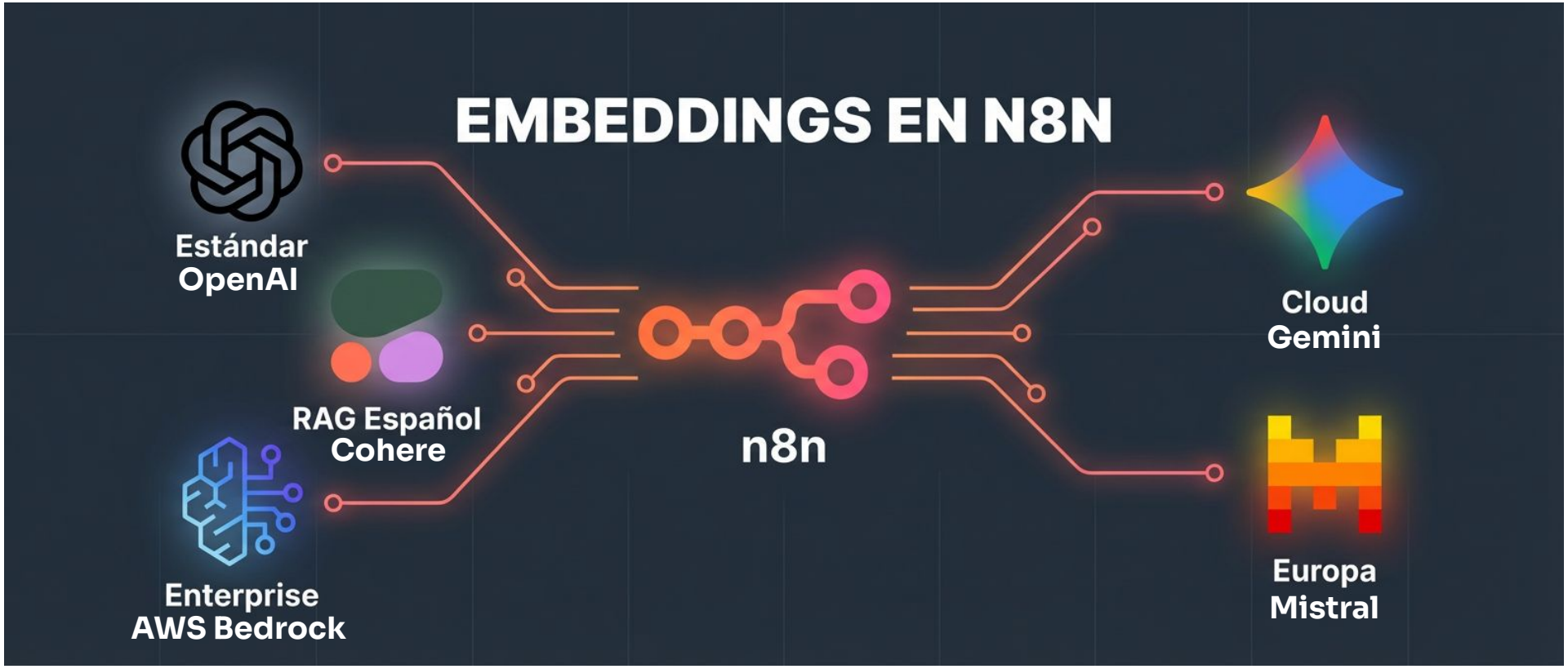
Similitud Semántica

"Me encanta el fútbol"	[0.89, 0.12, -0.34...]	BASE
"Adoro jugar al balón"	[0.87, 0.14, -0.31...]	✓ SIMILAR
"Hoy llueve mucho"	[-0.23, 0.78, 0.56...]	✕ DISTINTO

Nota: Los modelos de Embeddings hacen esto automáticamente con un solo clic. Tú no calculas nada.

MODELOS EN N8N

¿Modelos de Embeddings?



CONCEPTOS: EL MAPA

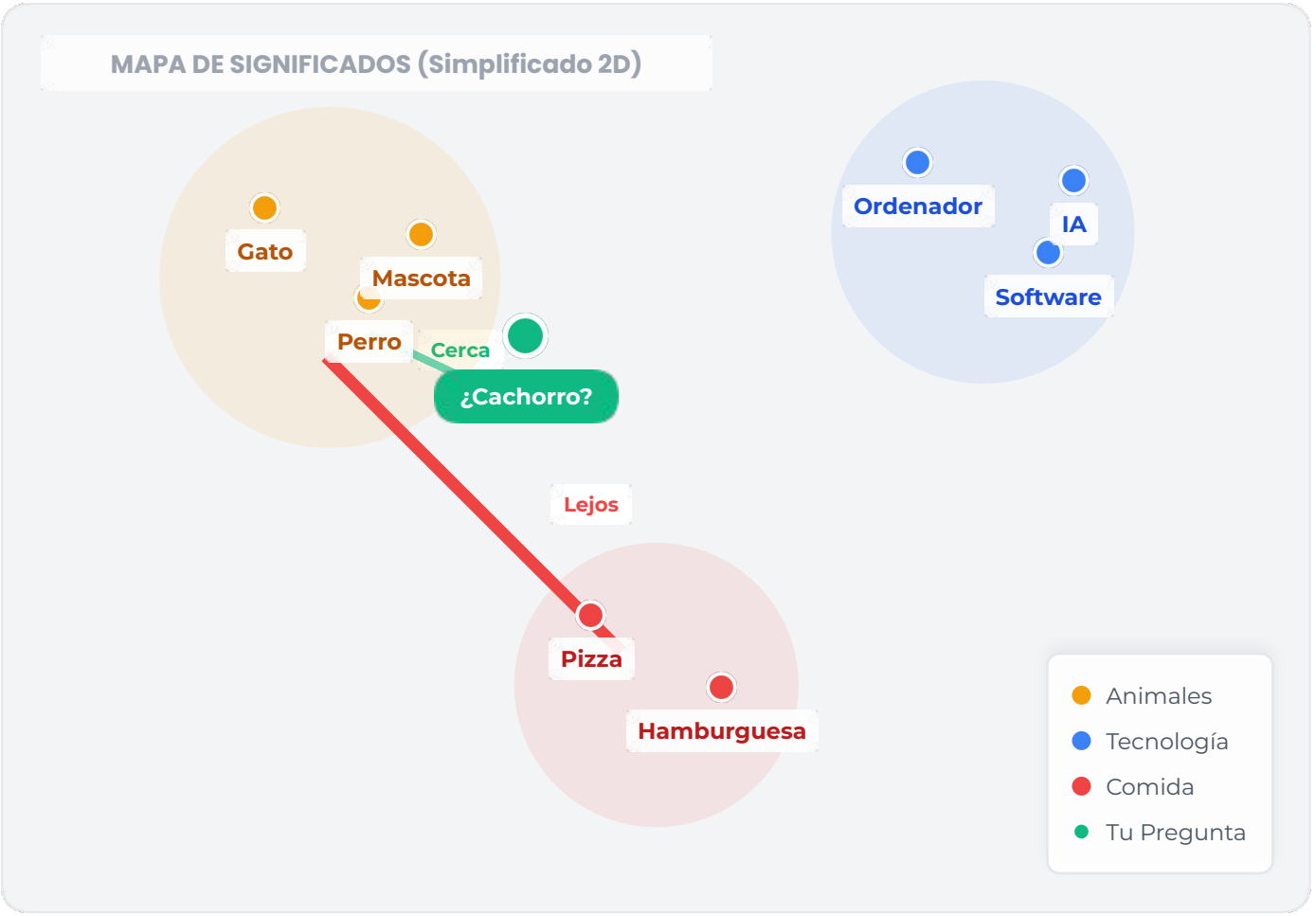
¿Qué es un Vector?

Definición Visual

Si el Embedding son las coordenadas, el **Vector** es el punto en el mapa.
Imagina un mapa gigante donde cada palabra o frase es un punto.

Analogía Google Maps

Cuando buscas "Restaurante italiano", Google Maps busca puntos **cercanos** a tu ubicación y filtra los lejanos.
La base de datos Vectorial hace lo mismo: busca los vectores (documentos) más cercanos a tu pregunta.



EL ALMACÉN DE VECTORES

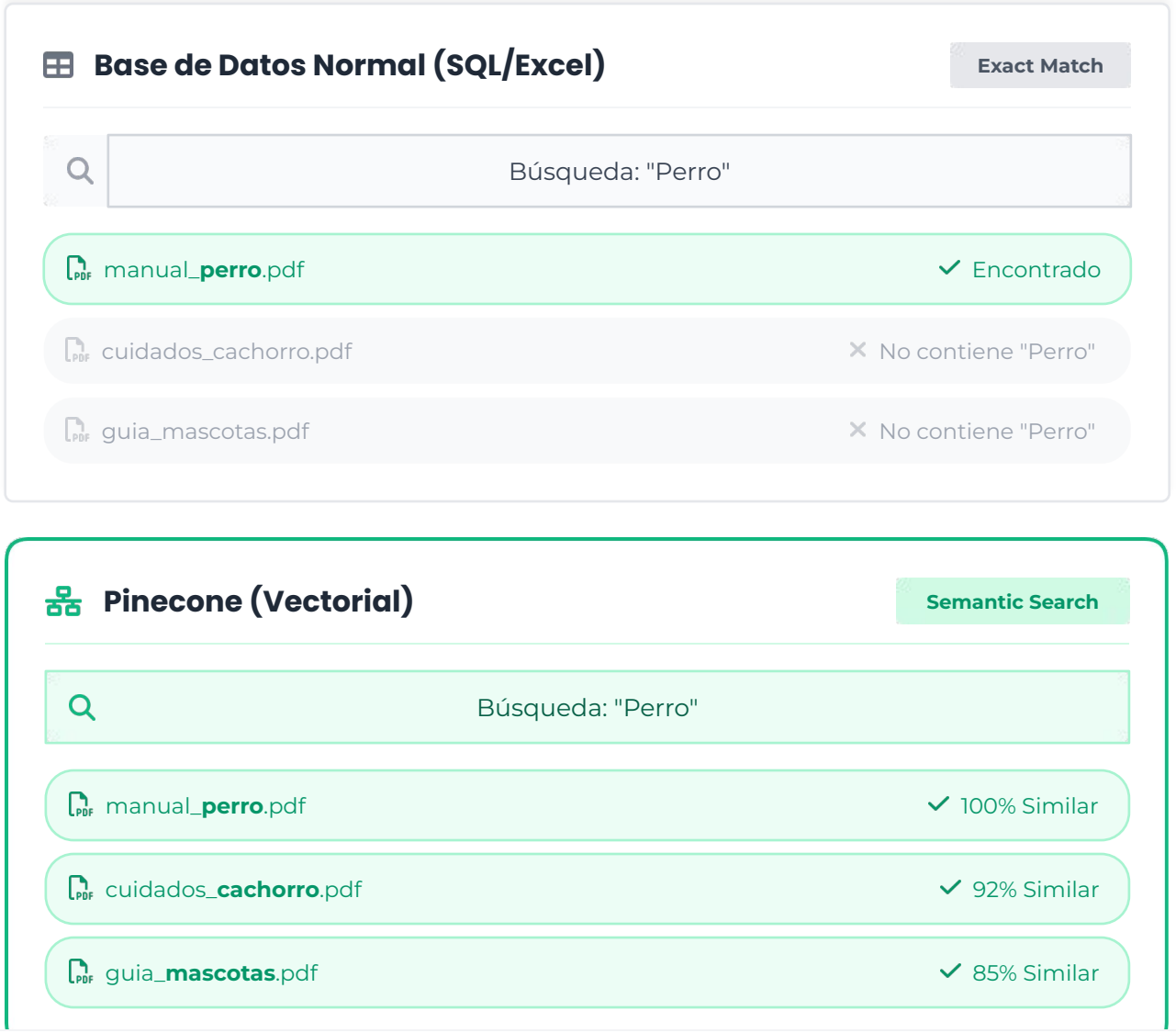
¿Qué es Pinecone?

Base de Datos Vectorial

Es un lugar diseñado específicamente para guardar y buscar vectores. No ordena por alfabeto ni por fecha, sino por **proximidad de significado**.

La Biblioteca Mágica

Imagina una bibliotecaria que ha leído TODOS los libros y los coloca por temas. Si pides "algo de cocina italiana", no busca la palabra exacta "cocina", sino que va a la sección donde están las recetas, la pasta y la pizza.



EL PROCESO

¿Qué es RAG?

Retrieval Augmented Generation

Es la técnica para conectar tu **"Cerebro"** (ChatGPT) con tu **"Memoria"** (Pinecone).

Analogía: El Examen con Chuleta

Imagina que ChatGPT va a un examen sobre tu empresa.

Sin RAG: Responde de memoria (se lo inventa).
Con RAG: Le dejamos sacar una chuleta (tus documentos) para copiar la respuesta exacta.

R — **RETRIEVAL (Recuperar)**
El usuario pregunta. Buscamos en Pinecone los fragmentos de documentos más relevantes (vectores cercanos).

A — **AUGMENTED (Aumentar)**
Cogemos esos fragmentos y los pegamos junto a la pregunta original. Creamos un "Prompt Enriquecido".

G — **GENERATION (Generar)**
Enviamos todo a ChatGPT. El modelo genera una respuesta usando SOLO la información que le acabamos de dar.

✕ Sin RAG

"Las devoluciones suelen ser de 30 días..." (Genérico/Alucinación)

✓ Con RAG

"Según vuestro manual, tenéis 15 días de devolución." (Dato Real)

RESUMEN

El Flujo Completo en 4 Pasos

FASE 1: PREPARACIÓN (Solo una vez)

Alimentar la Memoria

Se ejecuta una sola vez por documento

1. Cargar
PDFs, Docs, Web

2. Trocear
Chunks de texto

3. Embeddings
Texto a Números

4. Pinecone
Guardar Vectores

FASE 2: USO (Cada pregunta)

Consultar al Chatbot

Se ejecuta cada vez que el usuario pregunta

1. Pregunta
Usuario escribe

2. Embedding
Pregunta a Vector

3. Búsqueda
Encontrar Similar

4. Respuesta
ChatGPT + Contexto

CONFIGURACIÓN TÉCNICA

Configurar Pinecone

Prepara tu "biblioteca mágica" en 3 pasos.

1 Crear cuenta gratuita

Ve a **pinecone.io** y regístrate.

El plan "Starter" es gratis y permite hasta 100k vectores.

2 Crear un Index

Haz clic en "Create Index". Configura estos valores exactos:

Dimensiones: `1536`

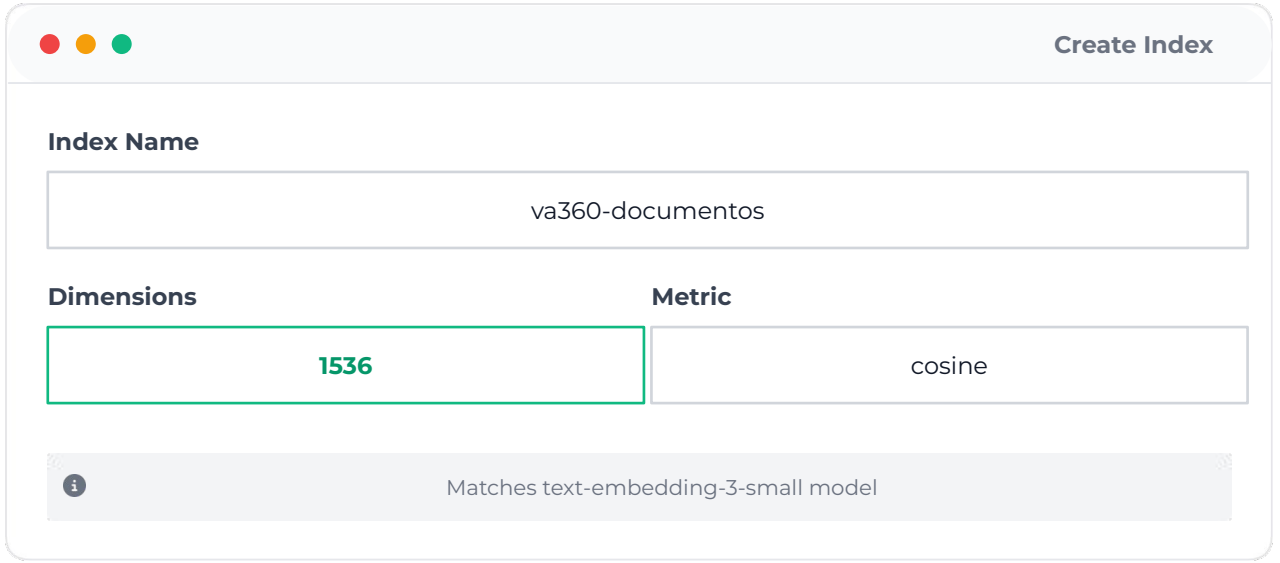
Métrica: `cosine`

1536 es clave para OpenAI.

3 Obtener Credenciales

En el menú lateral "API Keys", copia una cosa:

API Key: (Ej: pcl_823...)



CONECTANDO LAS PIEZAS

Configurar n8n

Conecta el "cerebro" (OpenAI) y la "biblioteca" (Pinecone).

1 Credencial de OpenAI

Ve a **Credenciales** → Añadir nueva → Busca "OpenAI".
Solo necesitas pegar tu **API Key** (la que empieza por `sk-...`).

Esto permite a n8n generar texto y embeddings.

2 Credencial de Pinecone

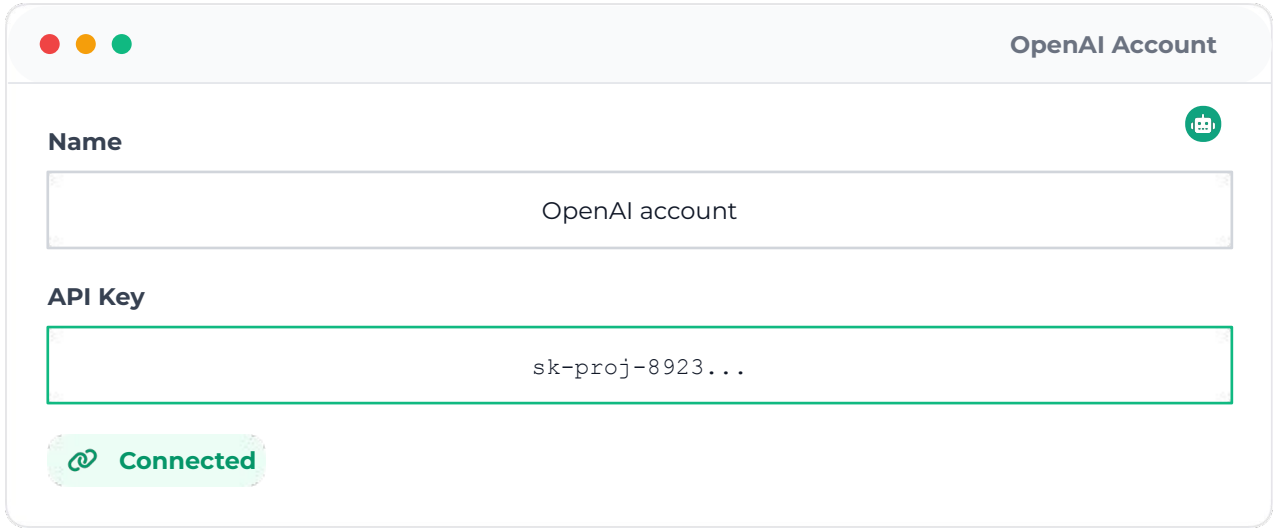
Busca "Pinecone" y rellena el campo clave:

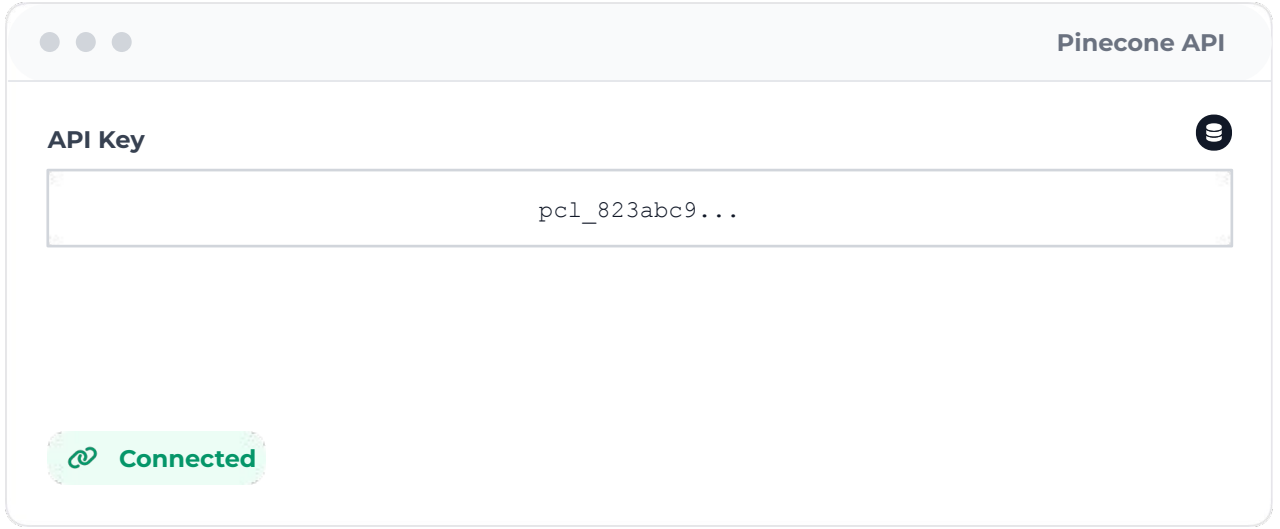
API Key: `pcl_823...`

3 Guardar y Verificar

Dale a **Save**. Si todo está bien, verás el estado "Connected".

¡Listo para usar en tus workflows!





Cargar Documentos en Pinecone

El proceso de "Ingestión": De archivo a vectores.

PREPARACIÓN DE DATOS

Chunking y Embeddings

4 reglas de oro para que tu chatbot sea inteligente.

Tamaño de Chunks Ideal

Divide el texto en trozos manejables.

`500-1500 caracteres` `10-20% solapamiento`

Limpieza de Texto

Elimina ruido que confunde al modelo: cabeceras, pies de página, números de página y saltos de línea extraños.

Modelo de Embeddings

Usa **text-embedding-3-small** (1536 dims) por defecto. Es barato y muy eficiente para la mayoría de casos.

Metadatos Inteligentes

No guardes solo texto. Añade contexto para filtrar mejor.

`source: manual.pdf` `category: rrhh`

Estrategia de Chunking

Carácteres por chunk	1000 chars
Solapamiento (Overlap)	200 chars

↑ El solapamiento mantiene el contexto entre cortes

Estructura del Vector

```
{
  "id": "vec_abc123",
  "values": [0.012, -0.234, 0.891, ...],
  "metadata": {
  "text": "La política de devoluciones..." ,
  "source": "manual_empleado.pdf" ,
  "page": 144,
  "date": "2023-10-15"
}
```

Chat con tus Datos (RAG)

El cerebro del sistema: Conectando preguntas con respuestas.

CONFIGURACIÓN & TESTING

AI Agent: Reglas y Pruebas

Define el cerebro y verifica que no "alucine".

1 System Prompt (El Cerebro)

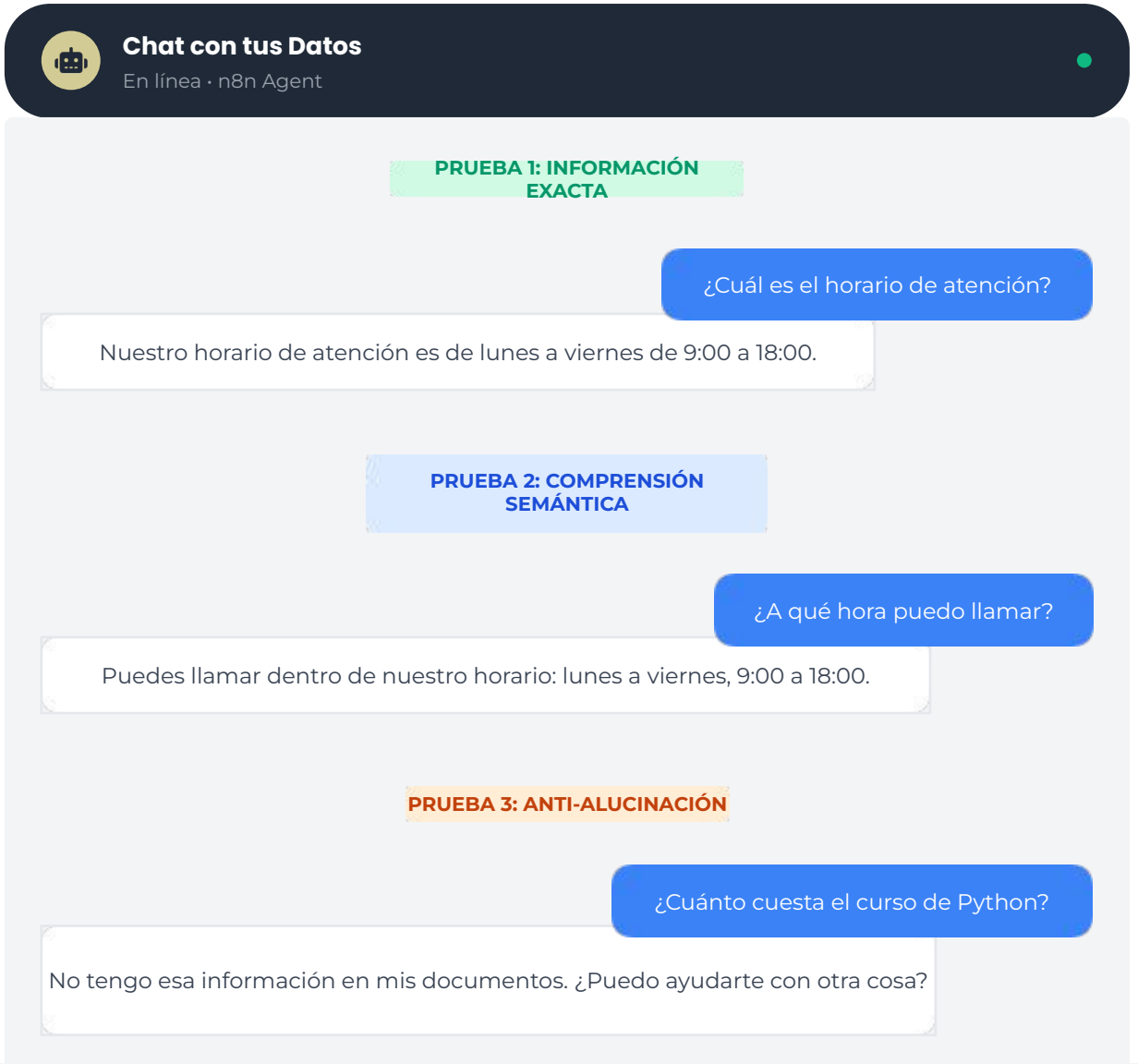
La instrucción más importante. Define límites claros para evitar que invente información.

```
Eres un asistente de VA360. Responde SOLO con la información
de los documentos proporcionados. Si no tienes la respuesta,
    di: "No tengo esa información en mis documentos".
```

2 Conectar Pinecone (La Memoria)

Configura la "Vector Store Tool" para que el agente pueda buscar.

Operation: Query Limit (Top K): 3-5



BUENAS PRÁCTICAS

Lo que SÍ debes hacer

Tamaño de Chunks Ajustado
Usa entre 500 y 1500 caracteres con solapamiento (10-20%) para mantener el contexto relevante.

Top K Moderado (3-5)
Recupera solo los fragmentos más relevantes para evitar ruido en la respuesta de la IA.

Prompt Anti-Alucinaciones
Instruye explícitamente: "Responde SOLO usando la información provista en el contexto".

Filtrado por Metadatos
Usa metadatos (año, dpto) para acotar la búsqueda antes de enviar vectores.

ERRORES COMUNES

Lo que debes EVITAR

Chunks Extremos
Muy pequeños pierden el sentido; muy grandes confunden al modelo con información basura.

Top K Excesivo (>10)
Enviar demasiados fragmentos diluye la información correcta y aumenta el coste.

Prompts Vagos
No restringir al modelo provoca que invente respuestas o use su conocimiento general.

Dimensiones Incorrectas
El índice (ej. 1536) debe coincidir exactamente con el modelo de embedding usado.

APLICACIÓN PRÁCTICA

Casos de Uso Reales

Esto no es teoría. Así es como las empresas están usando RAG hoy mismo.

Atención al Cliente 24/7

Chatbot que responde dudas sobre **FAQs, políticas de devolución y envíos** basándose en tus manuales.

Base de Conocimiento

Asistente interno para empleados: **RRHH, onboarding y soporte IT**. "¿Cómo pido vacaciones?".

Asistente Legal

Análisis rápido de **contratos y normativas**. Búsqueda de cláusulas específicas de penalización.

Asistente Comercial

Recomendaciones inteligentes de productos basadas en **catálogos técnicos** y necesidades del cliente.

Tutor Educativo

Tutor IA que responde dudas de los alumnos utilizando exclusivamente el **material del curso**.

Análisis Financiero

Extracción de datos clave y KPIs desde **informes trimestrales** y balances financieros complejos.

RESUMEN EN 30 SEGUNDOS

Embedding

Traductor de **palabras → números** que capturan el significado.

Vector

Un **punto** en el mapa semántico de tus conceptos.

Pinecone

El **mapa** (base de datos) donde guardas y buscas por significado.

RAG

Buscar trozos relevantes + IA respondiendo con **tus datos**.