



#GlobalAzure

#GABMUGPeru

#GABMUGPeru

Azure Open AI y Patron RAG

Carlos Augusto Diaz Huiza
Arquitecto Cloud



#GlobalAzure

Agenda

- Evolución AI
- Modelos LLM
- Conceptos Base
- Azure Open AI
- Patron RAG

#GABMUGPeru

Evolución IA



#GlobalAzure

Actualizaciones de Microsoft AI



Octubre 2022
Lanzamiento del
producto
Microsoft Designer



Enero 2023
Azure Open AI
Servicio disponible de
forma general



Marzo 2023
ChatGPT y GPT4
disponible en Azure Open AI
Service

Junio 2022

GitHub Copilot
disponible de
forma general



Enero 2023

Microsoft y OpenAI
amplían su asociación



Febrero 2023

Nuevo Bing y Edge
Con tecnología GPT4



Marzo 2023

Microsoft 365
Copilot,
Github introduce
Copilot X



#GABMUGPeru

Modelos LLM



#GlobalAzure

LLM (Large Language Models)

Modelos lingüísticos de gran tamaño de aprendizaje automático que pueden comprender y generar un texto en lenguaje humano. Funcionan al analizar conjuntos de datos masivos del lenguaje.



**Redes
neuronales**

**Simulando
escritura y habla
humana**

**Leer, traducir
y resumir
textos y crear
frases**

Azure OpenAI Service

GPT-4

GPT-4-Turbo

GPT-3.5-Turbo

New: GPT-4 for Vision

4K

8K

16K

32K

128K

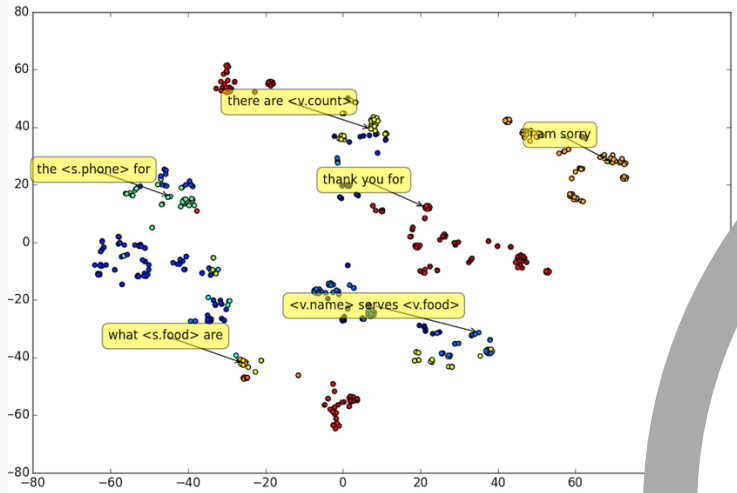
#GABMUGPeru

Conceptos Base



#GlobalAzure

Definiciones Importantes



Vectores

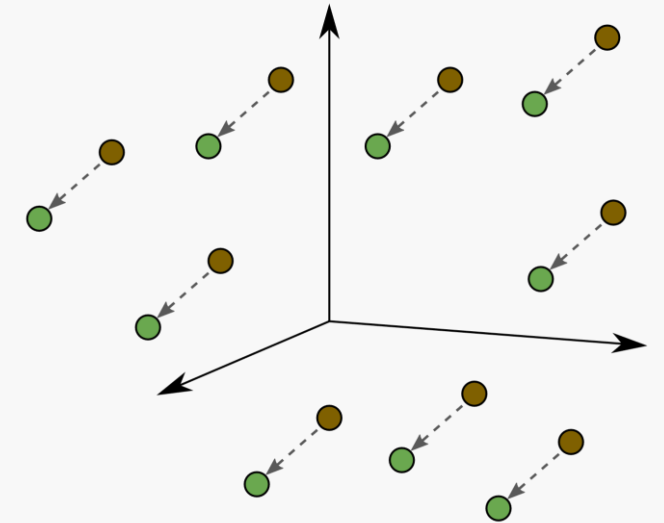
Un vector es un trazo que está conformado por miles de puntos.

Embedding

Técnica de procesamiento de lenguaje natural que convierte el lenguaje humano en vectores matemáticos

Token

Los tokens pueden ser palabras o meros fragmentos de caracteres



Tokenizer

GPT-3.5 & GPT-4 GPT-3 (Legacy)

hola como estas, quiero saber el estado de cuenta de mi tarjeta de crédito

Clear Show example

Tokens	Characters
19	73

hola como estas, quiero saber el estado de cuenta de mi tarjeta de crédito

- 4 caracteres continuos
- 1 palabra con hasta 4 caracteres
- Caracteres especiales

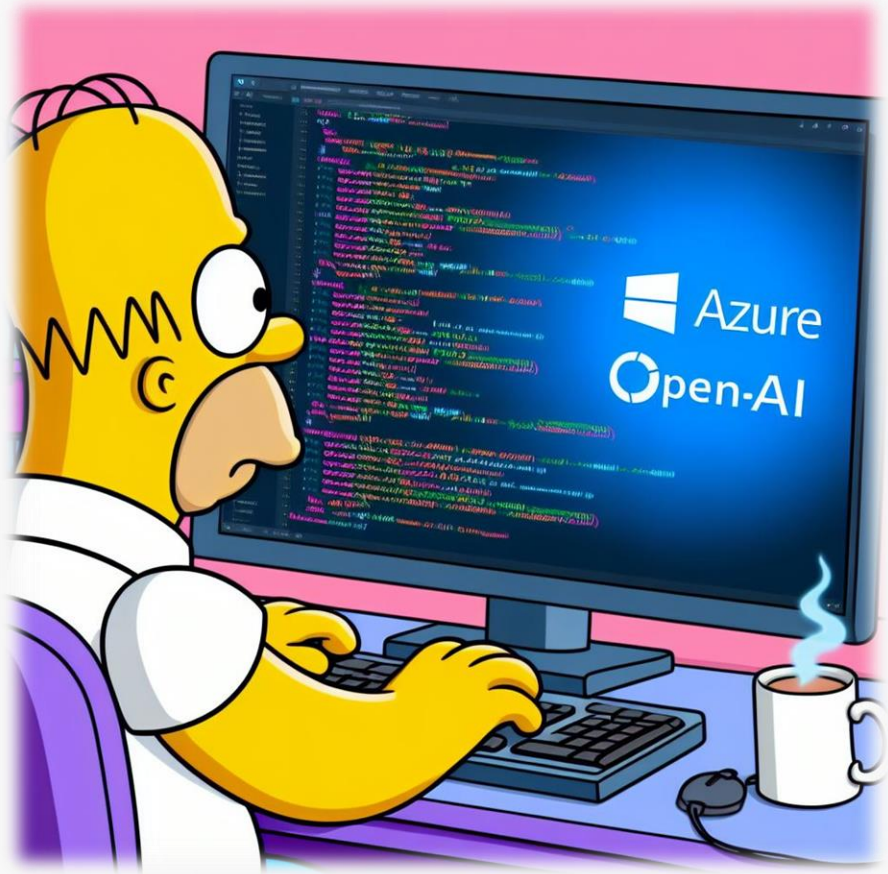
<https://platform.openai.com/tokenizer>

#GABMUGPeru

Que Azure Open AI



#GlobalAzure



Es una plataforma basada en la nube que permite a los desarrolladores y científicos de datos crear e implementar modelos de IA de forma rápida y sencilla

Azure AI Studio

Azure AI | Azure OpenAI Studio

 **Azure AI Studio** VERSIÓN PRELIMINAR PÚBLICA

Presentación del nuevo Azure AI Studio (versión preliminar)

Cree, evalúe e implemente sus soluciones de inteligencia artificial de un extremo a otro.

Explorar Inteligencia artificial de Azure Studio

Azure OpenAI

Área de juegos

Chat

Finalizaciones

DALL-E

Assistants (versión preliminar)

Administración

Implementaciones

Modelos

Archivos de datos

Cuotas


Azure OpenAI Studio

Le damos la bienvenida al servicio Azure OpenAI

Explore los modelos de inteligencia artificial generativos, cree solicitudes únicas para sus casos de uso y ajuste los modelos seleccionados.

Introducción

Generación de texto




Área de juegos de asistentes

VISTA PREVIA

Agilice el desarrollo de asistentes de inteligencia artificial con tecnología GPT con herramientas de personalización y administración del estado de conversación precompiladas.

Generación de texto



Traiga sus propios datos

Basa tus propios datos en modelos avanzados de inteligencia artificial para crear copilotos conversacionales que faciliten la comprensión del usuario, la finalización de tareas y la toma de decisiones.

Generación de texto



Área de juegos de chat

Diseñe un asistente de IA personalizado con ChatGPT. Experimente con los modelos GPT-3.5-Turbo y GPT-4.

Generación de texto



Área de juegos de finalizaciones

Experimente con modelos de finalizaciones para casos de uso como resumen, generación de contenido y clasificación.

Privacidad y cookies

Azure AI Studio

Azure AI Studio

VERSIÓN PRELIMINAR PÚBLICA

Presentación del nuevo Azure AI Studio (versión preliminar)

Cree, evalúe e implemente sus soluciones de inteligencia artificial de un extremo a otro.

Explorar Inteligencia artificial de Azure Studio

Azure OpenAI

Área de juegos

Chat

Finalizaciones

DALL-E

Assistants (versión preliminar)

Administración

Implementaciones

Modelos

Archivos de datos

Cuotas

Azure OpenAI Studio > Modelos

Privacidad y cookies

Modelos

Azure OpenAI cuenta con tecnología de modelos con diferentes funcionalidades y precios. Implemente uno de los modelos base proporcionados para probarlo en el área de juegos o entrene un modelo personalizado con sus datos y casos de uso específicos para mejorar el rendimiento y obtener resultados más precisos.
[Más información sobre los distintos tipos de modelos base](#)

Modelos base

Implementar

Crear un modelo personalizado


Opciones de columna

Actualizar

Buscar

Nombre del modelo ▾	Versión de m... ▾	Hora de creación: ▾	Estado ▾	Desplegable ▾
gpt-35-turbo	0125	14/2/2024 19:00	✓ Correcto	✓ Sí
gpt-35-turbo	1106	14/11/2023 19:00	✓ Correcto	✓ Sí
gpt-35-turbo	0613	18/6/2023 19:00	✓ Correcto	✓ Sí
gpt-35-turbo-16k	0613	18/6/2023 19:00	✓ Correcto	✓ Sí
gpt-4	0613	18/6/2023 19:00	✓ Correcto	✓ Sí

Azure AI Studio

 Azure AI Studio VERSIÓN PRELIMINAR PÚBLICA

Presentación del nuevo Azure AI Studio (versión preliminar)

Cree, evalúe e implemente sus soluciones de inteligencia artificial de un extremo a otro.

Explorar Inteligencia artificial de Azure Studio

Azure OpenAI

Área de juegos

Chat

Finalizaciones

DALL-E

Assistants (versión preliminar)

Administración

Implementaciones

Modelos

Archivos de datos

Cuotas

Azure OpenAI Studio > Área de juegos de chat

Área de juegos de chat

Importar configuración Exportar configuración Una nueva aplicación web...

Reproducir chat Borrar chat Configuración del área de juegos

Ver código ☐ Mostrar JSON

Uso de plantillas

Use una plantilla para empezar o simplemente empiece a escribir su propio mensaje del sistema a continuación. ¿Quiere algunas sugerencias? [Más información](#)

Seleccionar una plantilla

Mensaje del sistema

You are an AI assistant that helps

Pruebe el asistente mediante el envío de consultas a continuación. A continuación, ajusta la configuración del asistente para

Escriba aquí la consulta de usuario. (Mayús + Entrar para una nueva línea)

Configuración

Implementación Parámetros

Implementación *

gptturbotest

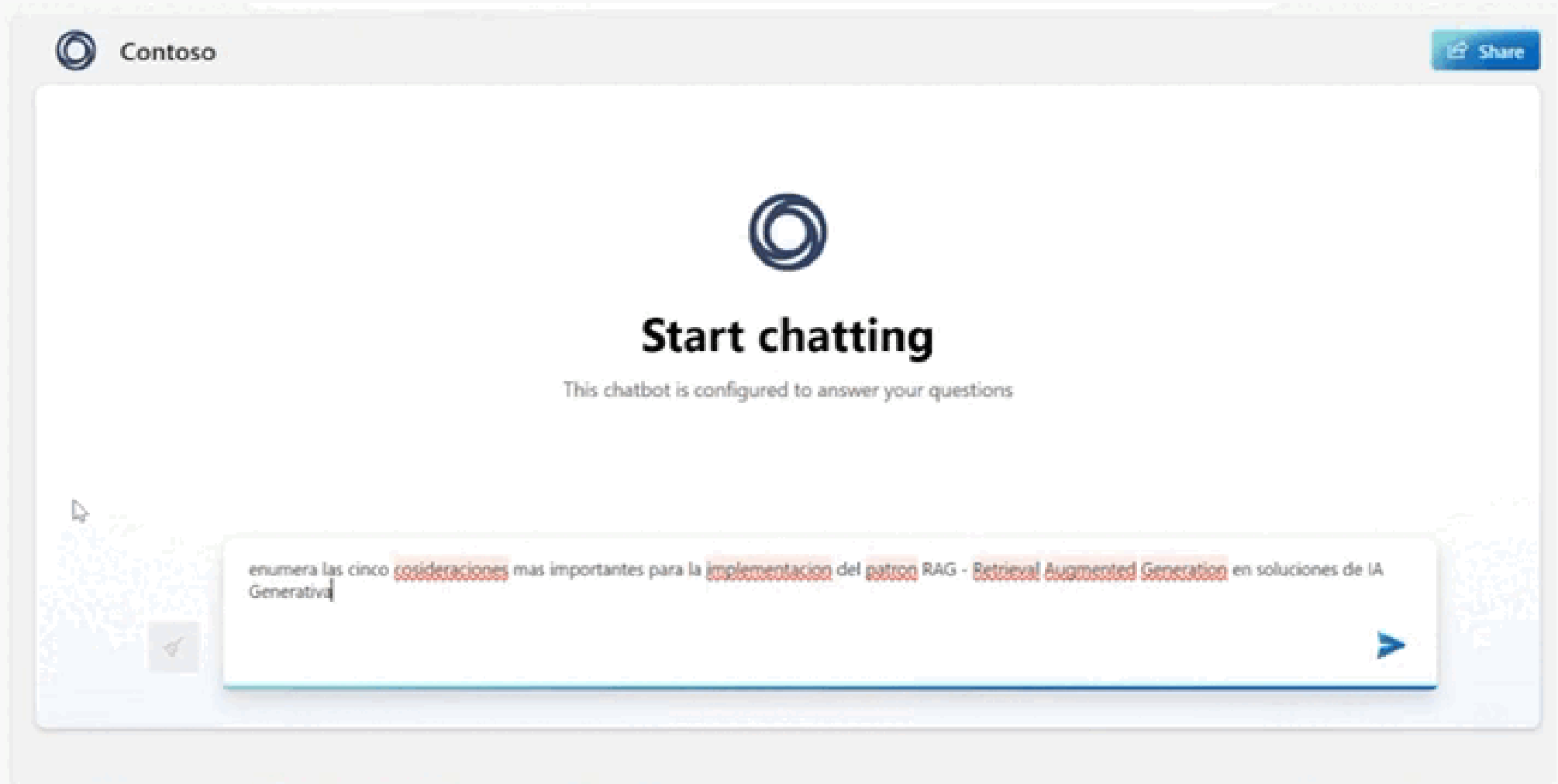
Configuración de sesión

Mensajes anteriores incluidos

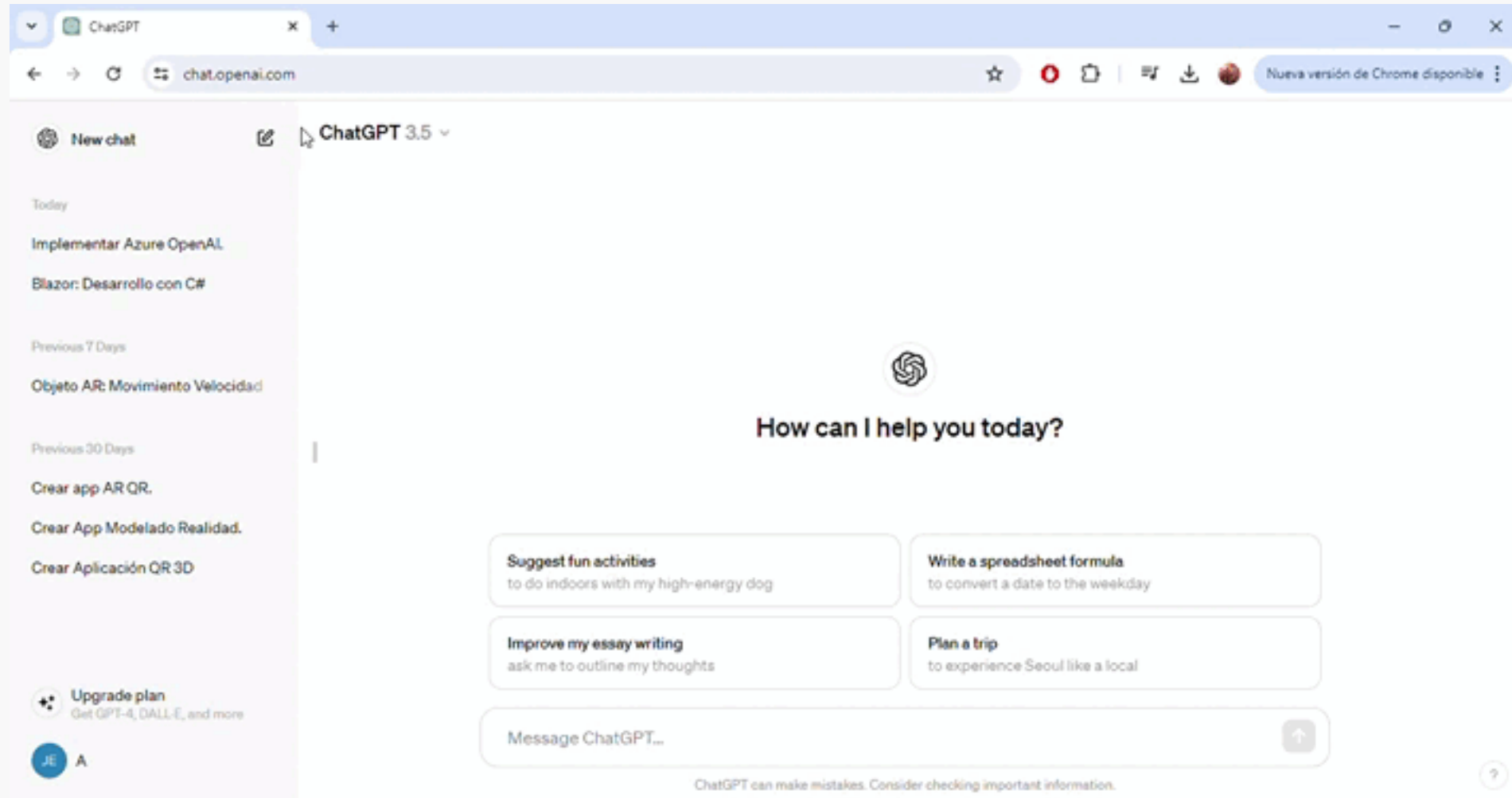
10

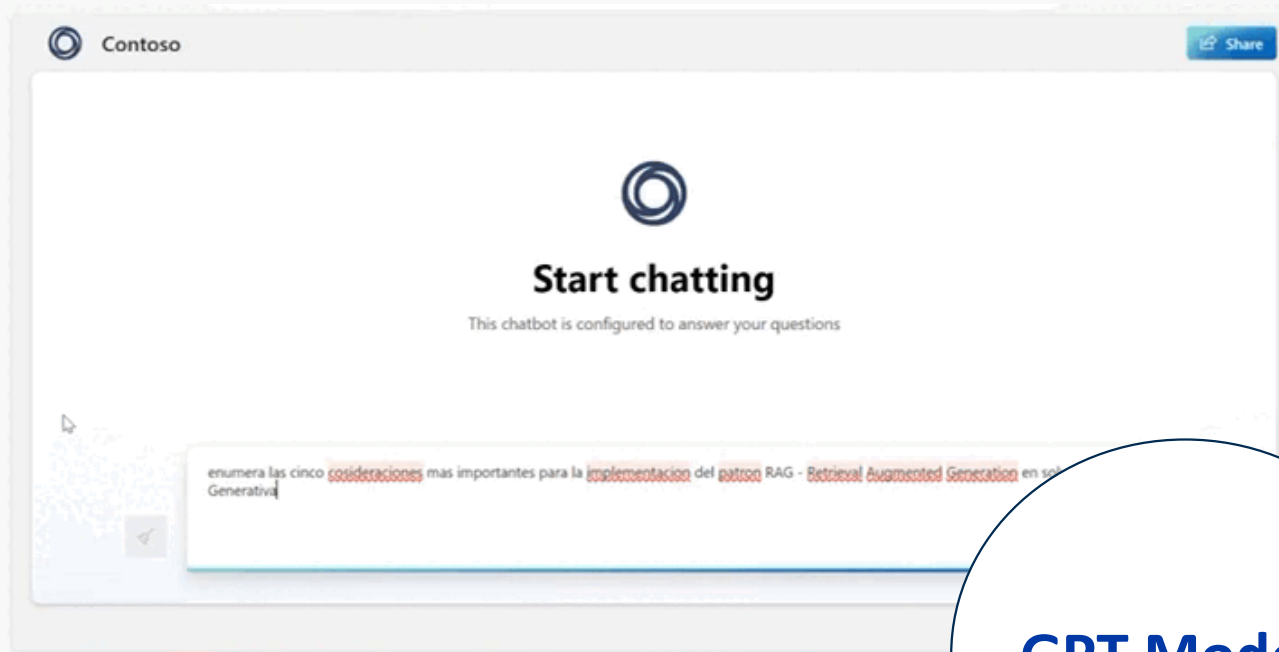
Recuento de tokens actual

Azure AI Studio

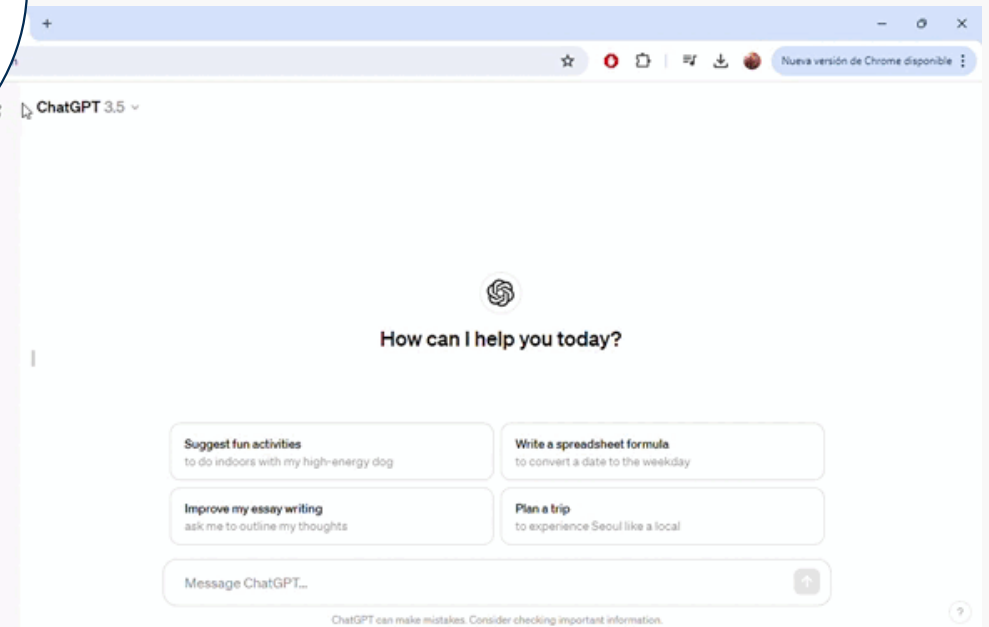


Open AI - ChatGPT





GPT Models



Pricing Options

Modelos de Lenguaje

Models	Context	Input (Per 1,000 tokens)	Output (Per 1,000 tokens)
GPT-3.5-Turbo-0125	16K	\$0.0005	\$0.0015
GPT-3.5-Turbo-Instruct	4K	N/A	N/A
GPT-4-Turbo	128K	\$0.01	\$0.03
GPT-4-Turbo-Vision	128K	\$0.01	\$0.03
GPT-4	8K	\$0.03	\$0.06
GPT-4	32K	\$0.06	\$0.12

Modelo de Embedding

Models	Per 1,000 tokens
Ada	\$0.0001
text-embedding-3-large	\$0.00013
text-embedding-3-small	\$0.00002

Modelo de Habla

Models	Price
Whisper	\$0.36/hour
TTS (Text to Speech)	\$15/1M characters
TTS HD	\$30/1M characters

Modelos Básicos

Modelos	Uso por 1000 tokens
Babbage-002	\$0,0004
Davinci-002	\$0,002

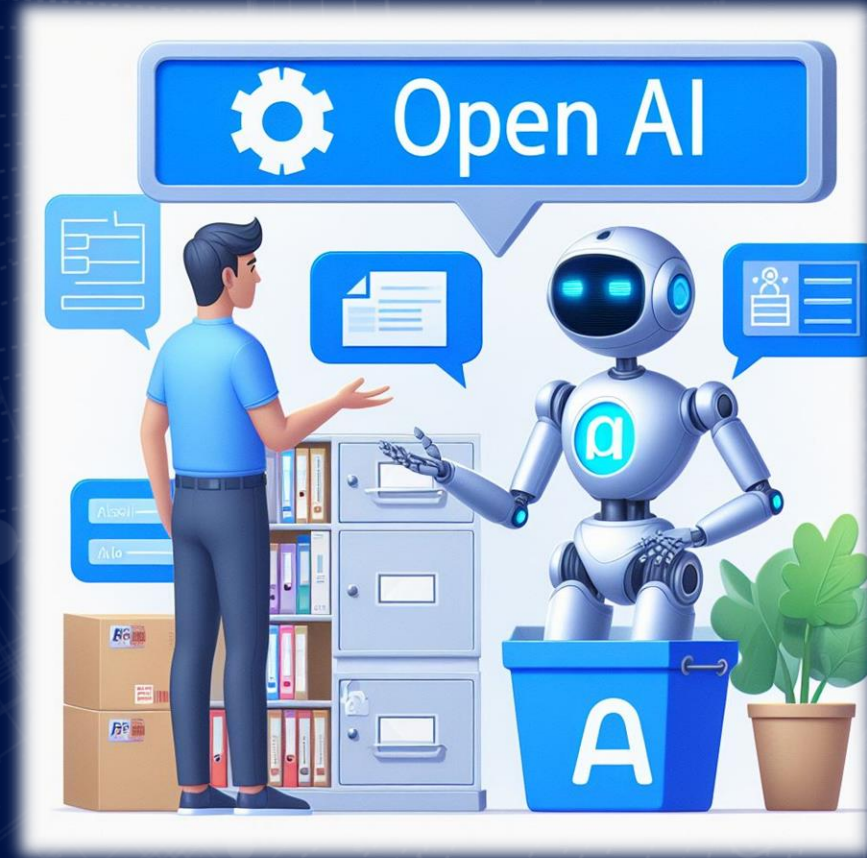
Modelos para Fine-tuning

Models	Training per compute hour	Hosting per hour	Input Usage per 1,000 tokens	Output Usage per 1,000 tokens
Babbage-002	N/A	N/A	N/A	N/A
Davinci-002	N/A	N/A	N/A	N/A
GPT-3.5-Turbo (4K)	\$45	\$3	\$0.0005	\$0.0015
GPT-3.5-Turbo (16K)	\$68	\$3	\$0.0005	\$0.0015

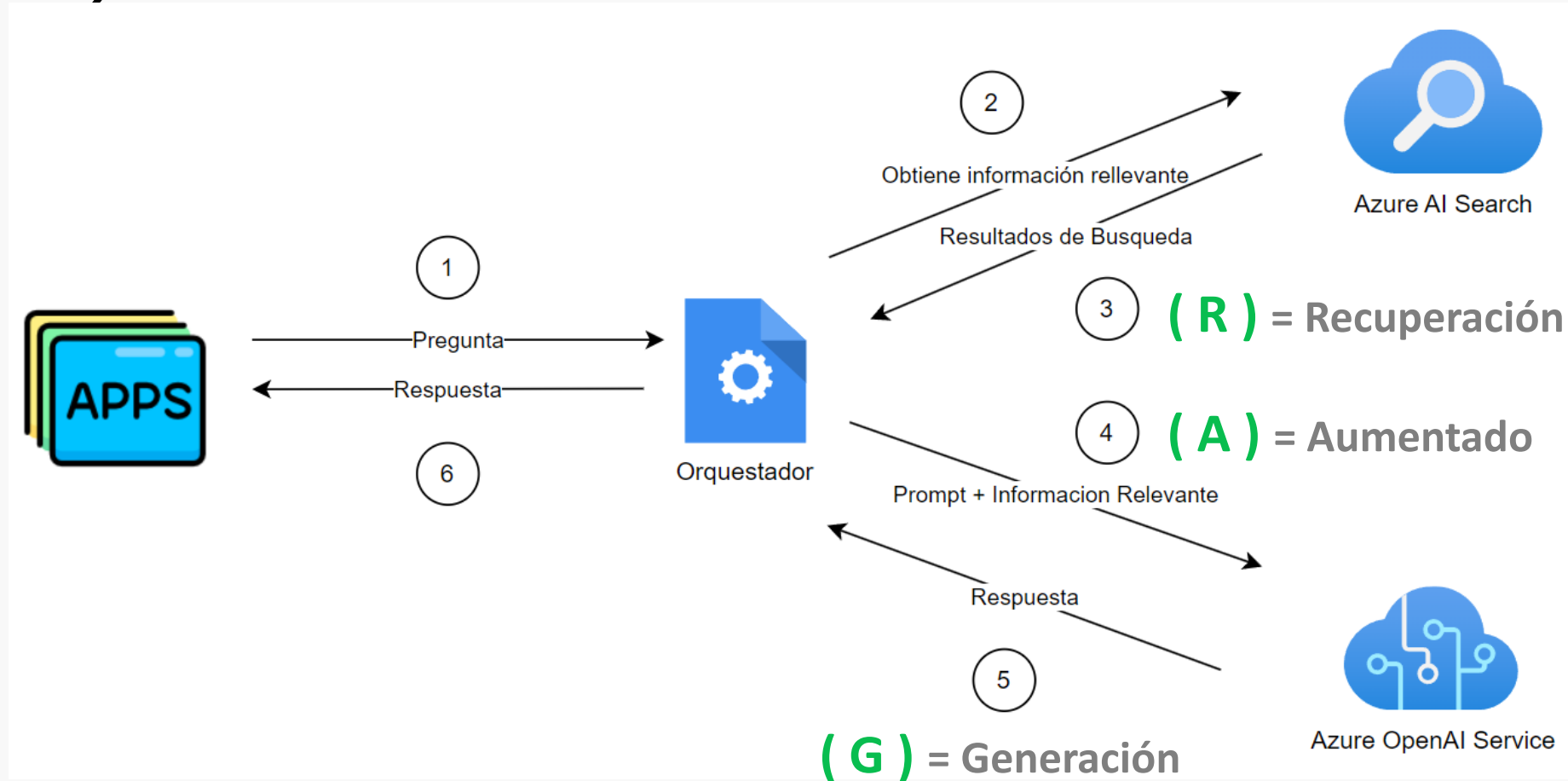
Modelo de Imágenes

Models	Quality	Resolution	Price (per 100 images)
Dall-E-3	Standard	1024 * 1024	\$4
	Standard	1024 * 1792, 1792 * 1024	\$8
Dall-E-3	HD	1024 * 1024	\$8
	HD	1024 * 1792, 1792 * 1024	\$12
Dall-E-2	Standard	1024 * 1024	\$2

Patron RAG

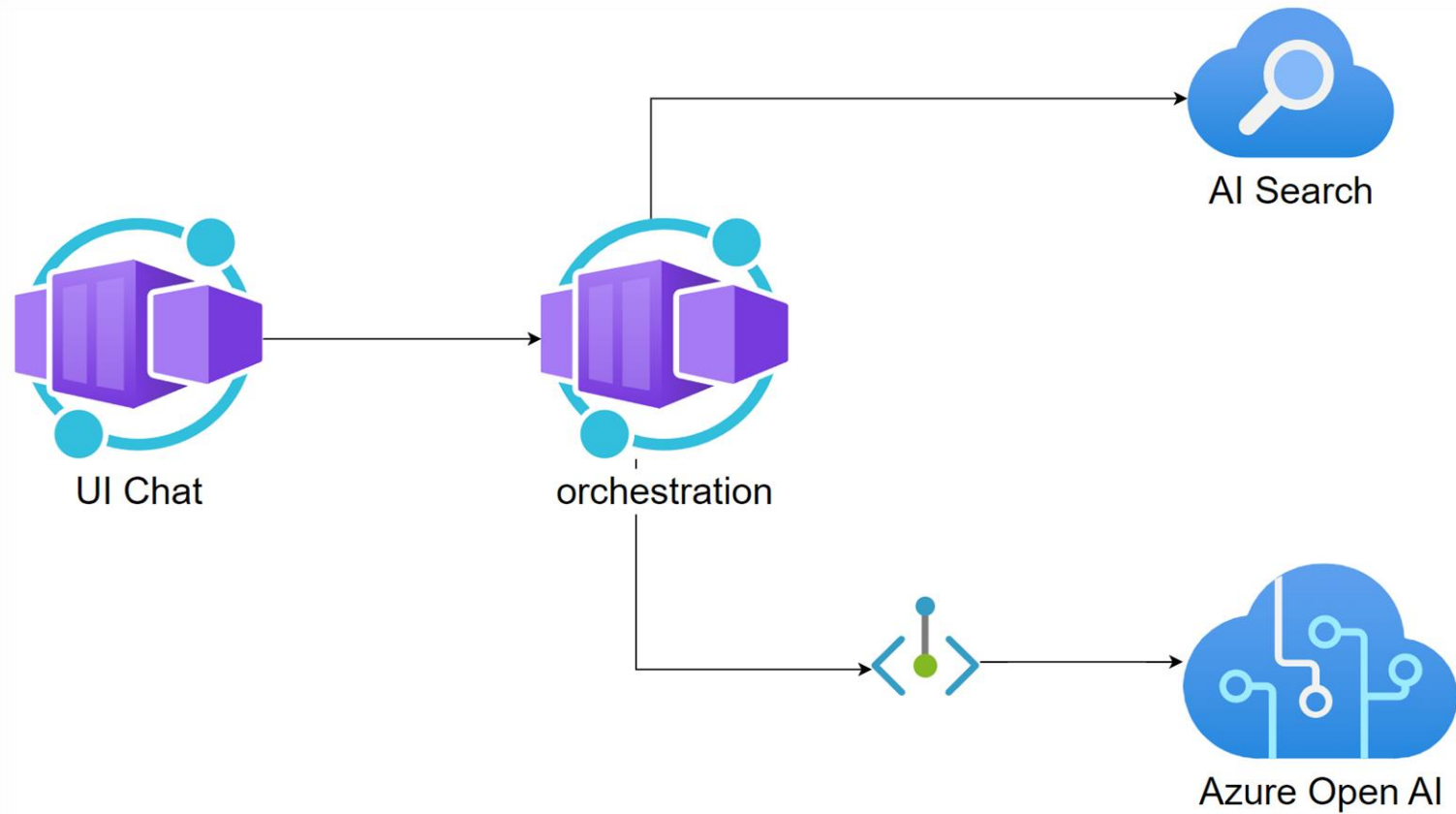


Patrón Azure Open AI con Base de Referencia (RAG)

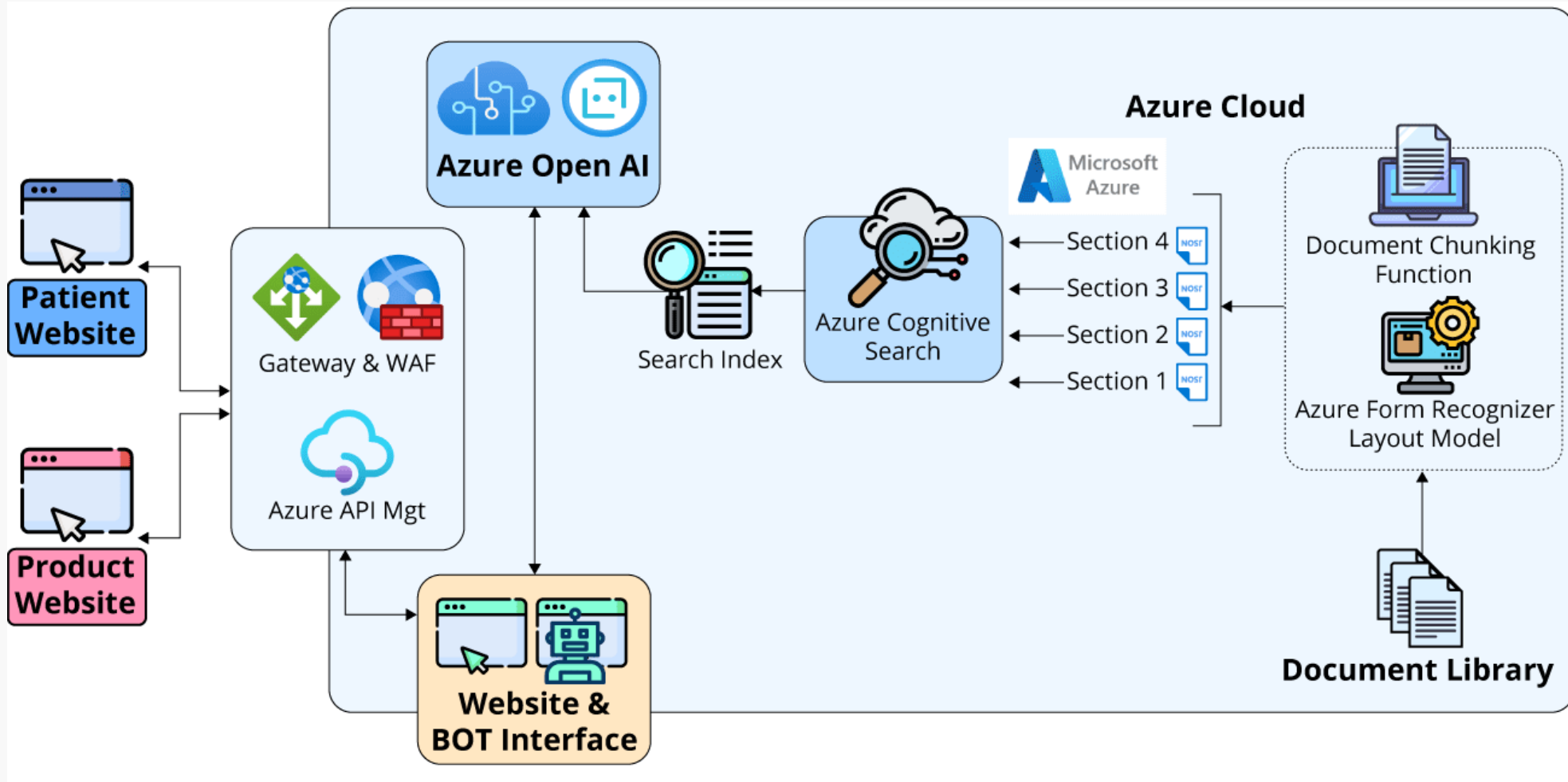


Retrieval Augmented Generation (RAG)

Patrón Azure Open AI con Base de Referencia (RAG)



Soluciones



Caso de Estimación

PROMT

Eres un asesor de ventas con experiencia en seguros de salud

Tokens: 15

Pregunta

Información Relevante

Respuesta

Hola, que tal

Me podrías indicar que seguro recomiendas para una persona de 50 años que suele atenderse en una clínica internacional por revisiones de rodilla

Tokens: 42

1. Cobertura Integral de Salud

Tokens: 459

Recomendaría el Plan Integral de que ofrece una completa para consultas y gastos hospitalarios. Proporciona acceso a una.

Tokens: 155

Costo Total por Consulta = 0.0010882

Embedding ADA

$(42+15) \times 0.0000001 = 0.0000042$

Solicitar GPT-3.5-Turbo

$(42+15+459) \times 0.0000015 = 0.000774$

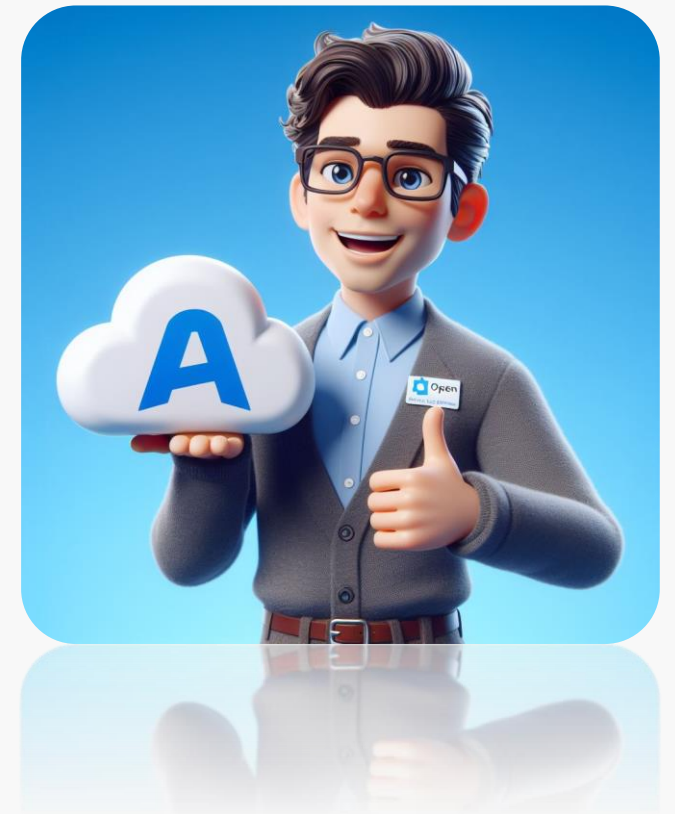
Respuesta GPT-3.5-Turbo

$155 \times 0.000002 = 0.00031$

Next steps

- Documentation <https://learn.microsoft.com/es-es/azure/ai-services/openai/how-to/create-resource?pivots=web-portal>

Todas las imágenes fueron generadas por Dall-E



Patrocinadores



Microsoft



mibanco



Colabora



nuvem

inetum.

Positive digital flow

