

# Desplegando aplicaciones de Large Language Models (LLMs) en la nube de Microsoft Azure.

Noviembre 25, 2023, a las 3:00 P.M.



**Kevin Knights**

Machine Learning Engineer



# Hola, soy Kevin Knights!

## Machine Learning Engineer

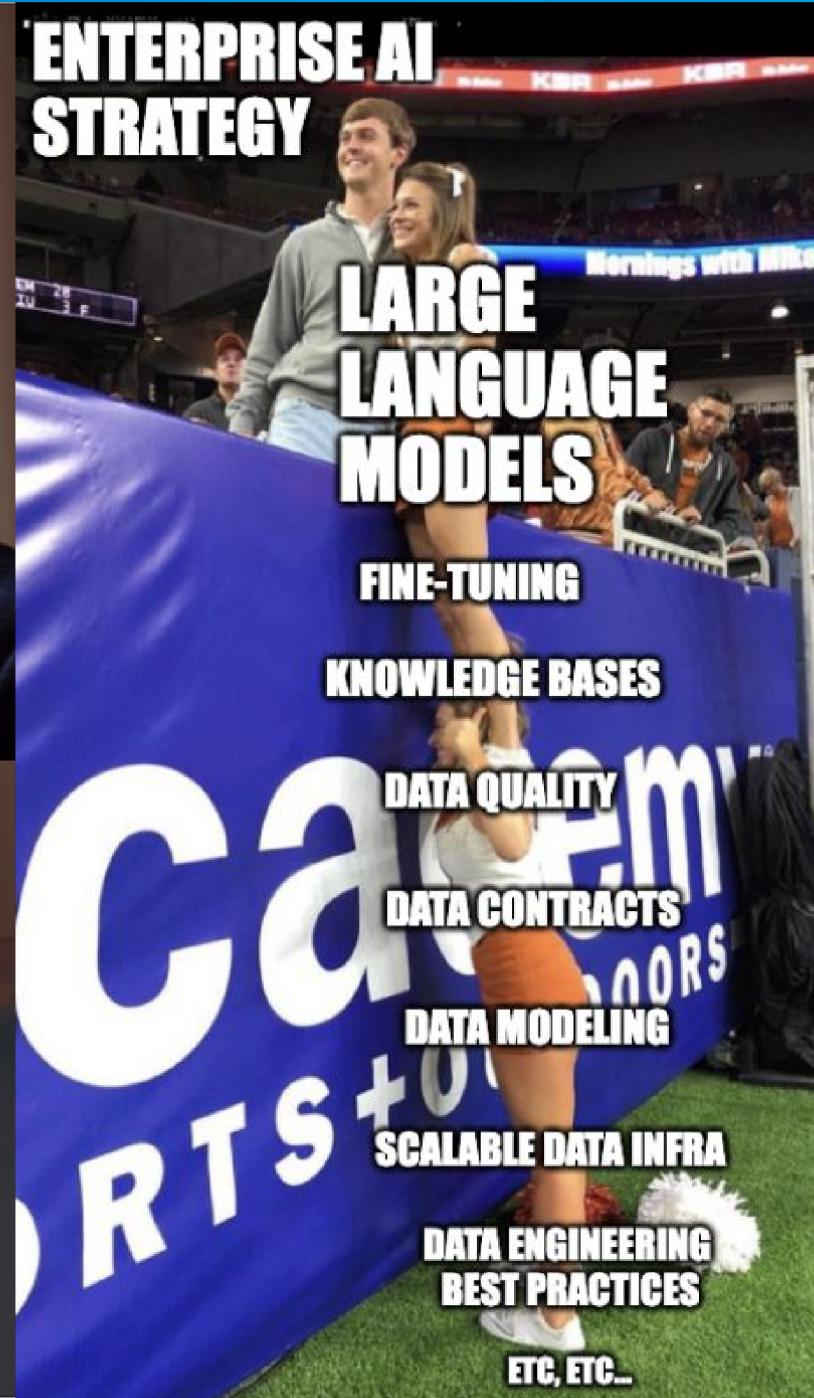
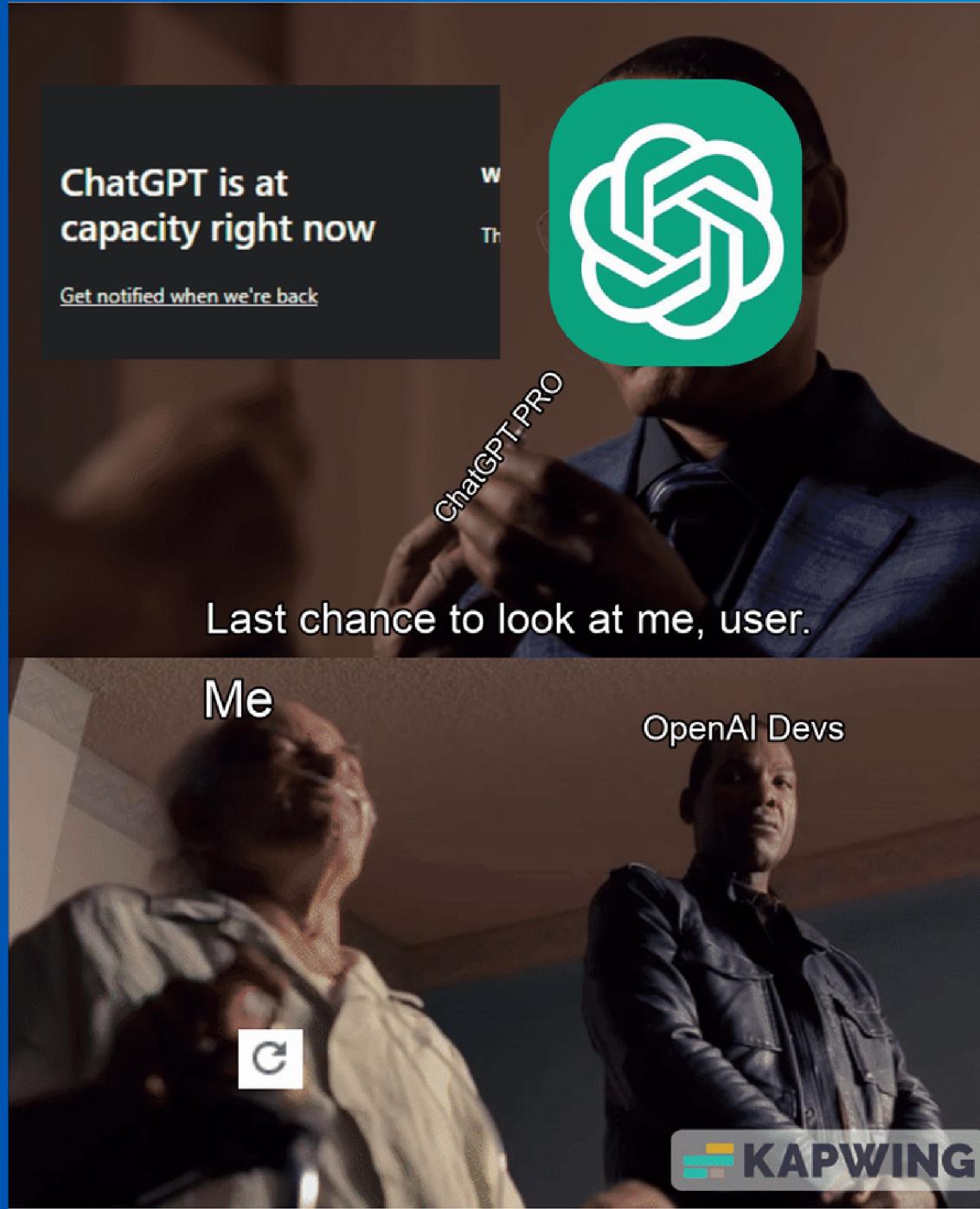
- He desarrollado varios productos GenAI.
- He participado en conferencias globales sobre IA.
- Certificado en AI de Stanford, Microsoft, entre otros...

LinkedIn: Kevin Knights

<https://www.linkedin.com/in/knightsk/>



# ¡Todo ahora es con LLMs (GenAI)!



Google dijo 'AI' más de 140 veces en su discurso de apertura del Google I/O de 2 horas

# ¿Qué es GenAI?

La inteligencia artificial generativa (GenAI) describe algoritmos (como ChatGPT) que se pueden utilizar para crear contenido nuevo, incluido audio, código, imágenes, texto, simulaciones y videos.

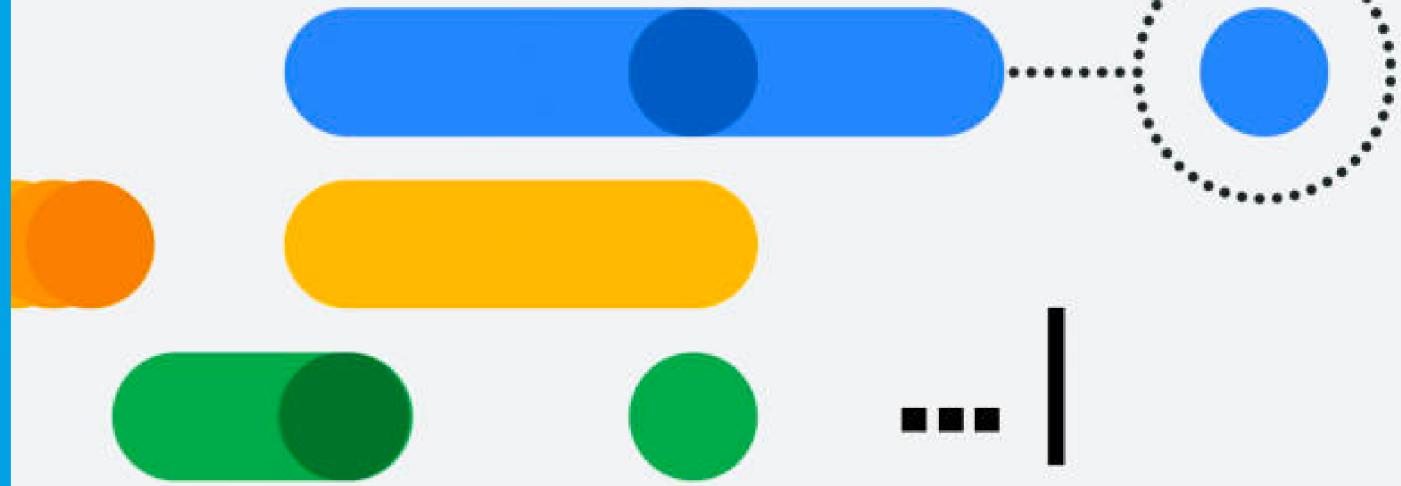
Fuente: [McKinsey – What is generative AI?](#)



**ChatGPT**

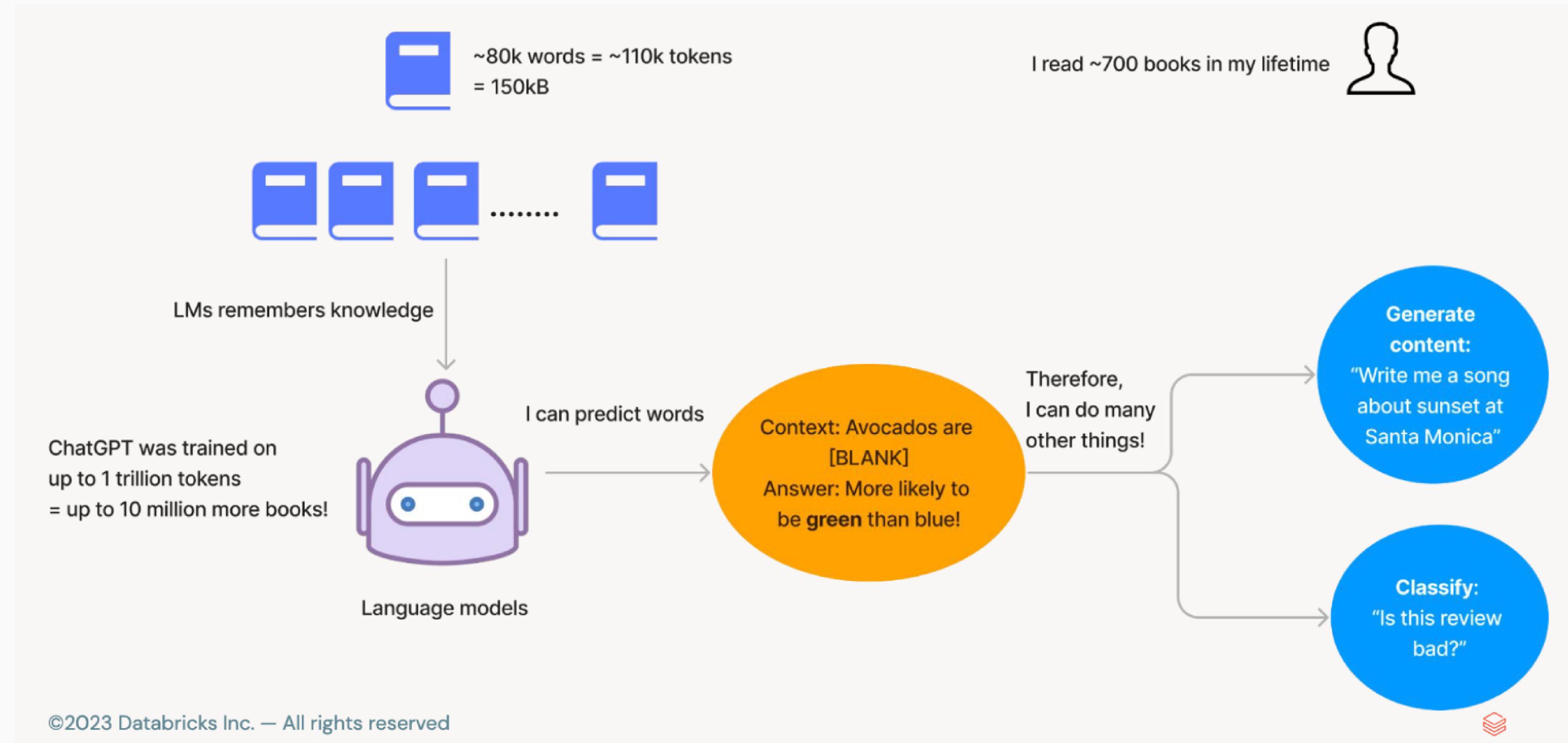


**Bard AI**



# ¿Qué son los Large Language Models?

Los modelos de lenguaje grande (LLM) son modelos de GenAI que pueden comprender y generar texto en lenguaje humano.



# ¿Por qué implementar LLMs?

McKinsey estima que GenAI (LLMs) agregará entre **2,6 y 4,4 billones de dólares** en valor anual a la economía global.

40%

El nuevo impacto económico de la IA como un todo, gracias a GenAI (LLMs).

No puedo pensar en nada que haya sido más poderoso desde la computadora de escritorio. – Michael Carbin, MIT, MosaicML.

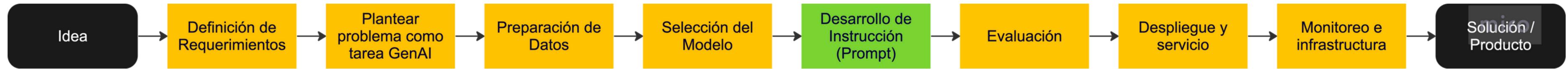
# LLMs y la transformación digital



Github Copilot, herramienta de generación de código

- GenAI y los LLM están democratizando el acceso a la IA, provocando los inicios de una IA verdaderamente empresarial.
- Ahora, una gran cantidad de datos no estructurados, pueden ser aprovechados para generar valor empresarial.
- Los LLMs liberarán a la fuerza laboral del trabajo que requiere mucho tiempo, para centrarse en áreas de conocimiento, estrategia y tareas de alto valor comercial.

# ¿Cómo construir aplicaciones con LLMs?



Marco para construir soluciones de GenAI.

- 1.Definición de requerimientos.
- 2.Plantar el problema como una tarea de ML.
- 3.Preparación de datos.
- 4.Selección del modelo.
- 5.Desarrollo de instrucción (prompt)
- 6.Evaluación.
- 7.Despliegue y servicio.
- 8.Monitoreo e infraestructura.



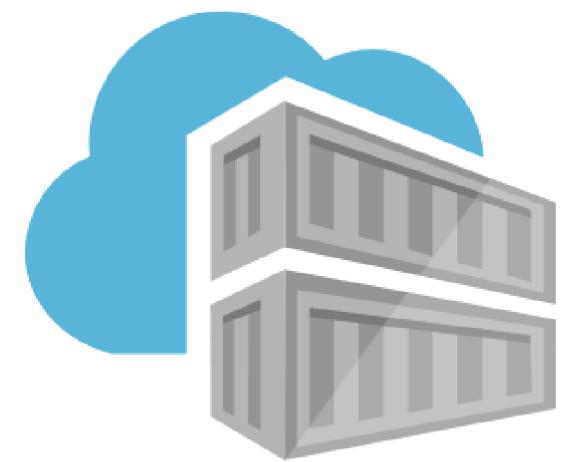
Python



MS Azure



# ¿Cómo construir LLMs apps en Azure?



Azure Container Registry

Administrar imágenes de Docker



Azure Kubernetes Service(AKS)

Despliegue de aplicaciones  
contenerizadas con k8s



Azure OpenAI

Servicios y APIs de LLMs

# Construyendo una solución de GenAI

## Idea:

Aplicación para generar contenido de redes sociales.

## Objetivo de Negocio:

Posicionar productos de forma personalizada.

## Requerimientos:

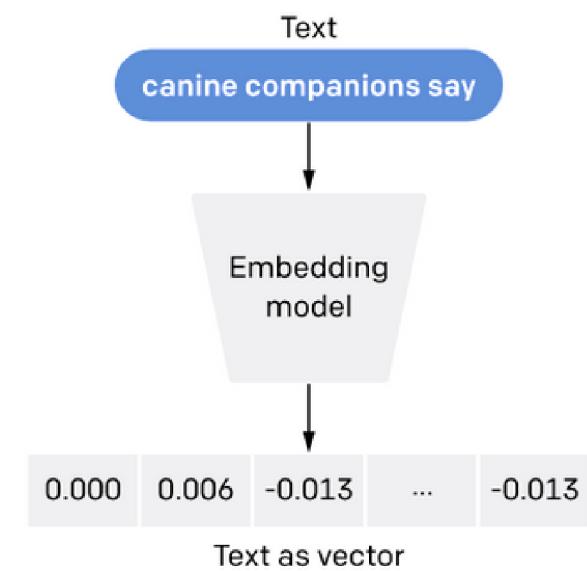
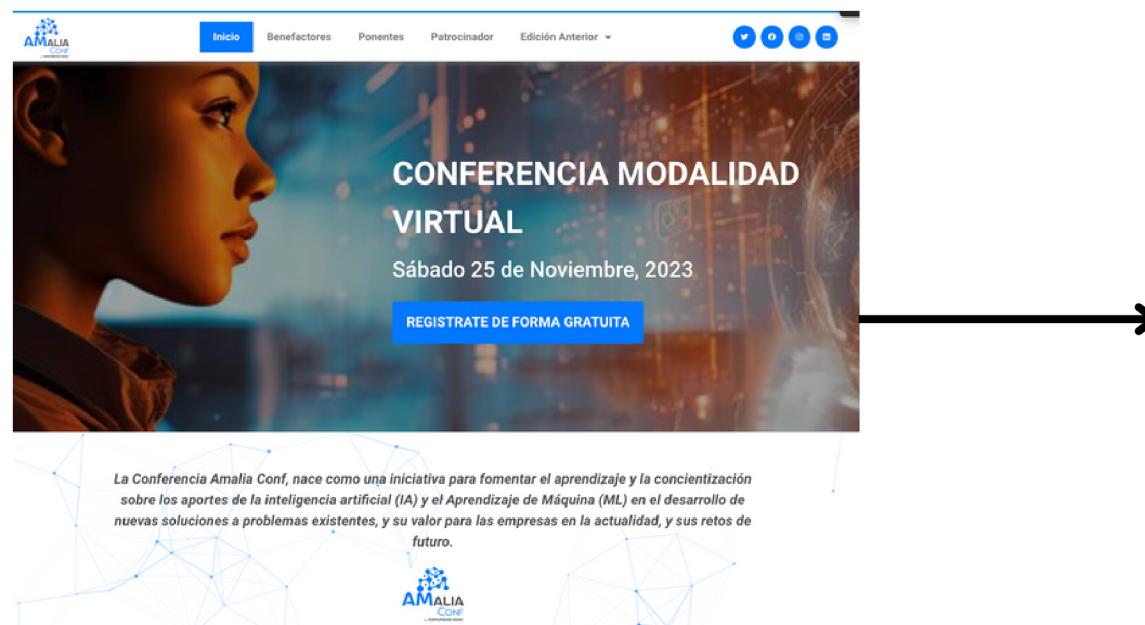
- Interfaz.
- Subir documentos.
- Personalizar contenido.

## Objetivo de LLM:

Maximizar la personalización del contenido.

# Construyendo una solución con LLMs

## Preparación de datos:



# Construyendo una solución con LLMs

The screenshot shows the Azure AI Studio interface, specifically the 'Models' section. On the left, there's a sidebar with links like 'Azure OpenAI', 'Playground', 'Chat', 'Completions', 'DALL-E (Preview)', 'Management', 'Deployments', 'Models' (which is currently selected), 'Data files', 'Quotas', and 'Content filters (Preview)'. The main area is titled 'Models' and contains a sub-section 'Base models'. It features a table with columns: Model name, Model version, Created at, Status, and Deployable. The table lists several models:

Model name	Model version	Created at	Status	Deployable
gpt-35-turbo	0613	6/18/2023 5:00 PM	Succeeded	Yes
gpt-35-turbo	0301	3/8/2023 4:00 PM	Succeeded	Yes
gpt-35-turbo-16k	0613	6/18/2023 5:00 PM	Succeeded	Yes
text-embedding-ada-002	2	4/2/2023 5:00 PM	Succeeded	Yes
text-embedding-ada-002	1	2/1/2023 4:00 PM	Succeeded	Yes

## Selección del modelo:

gpt3.5-turbo de OpenAI [1]  
text-embedding-ada-002 [2]

- Modelo con alta precisión.
- Buena calidad/precio.
  - ~ 750 palabras = 0.002 USD [1]
  - ~750 palabras = 0.0001 USD [2]
- Fácil integración

Precios de Open AI API: <https://openai.com/pricing>

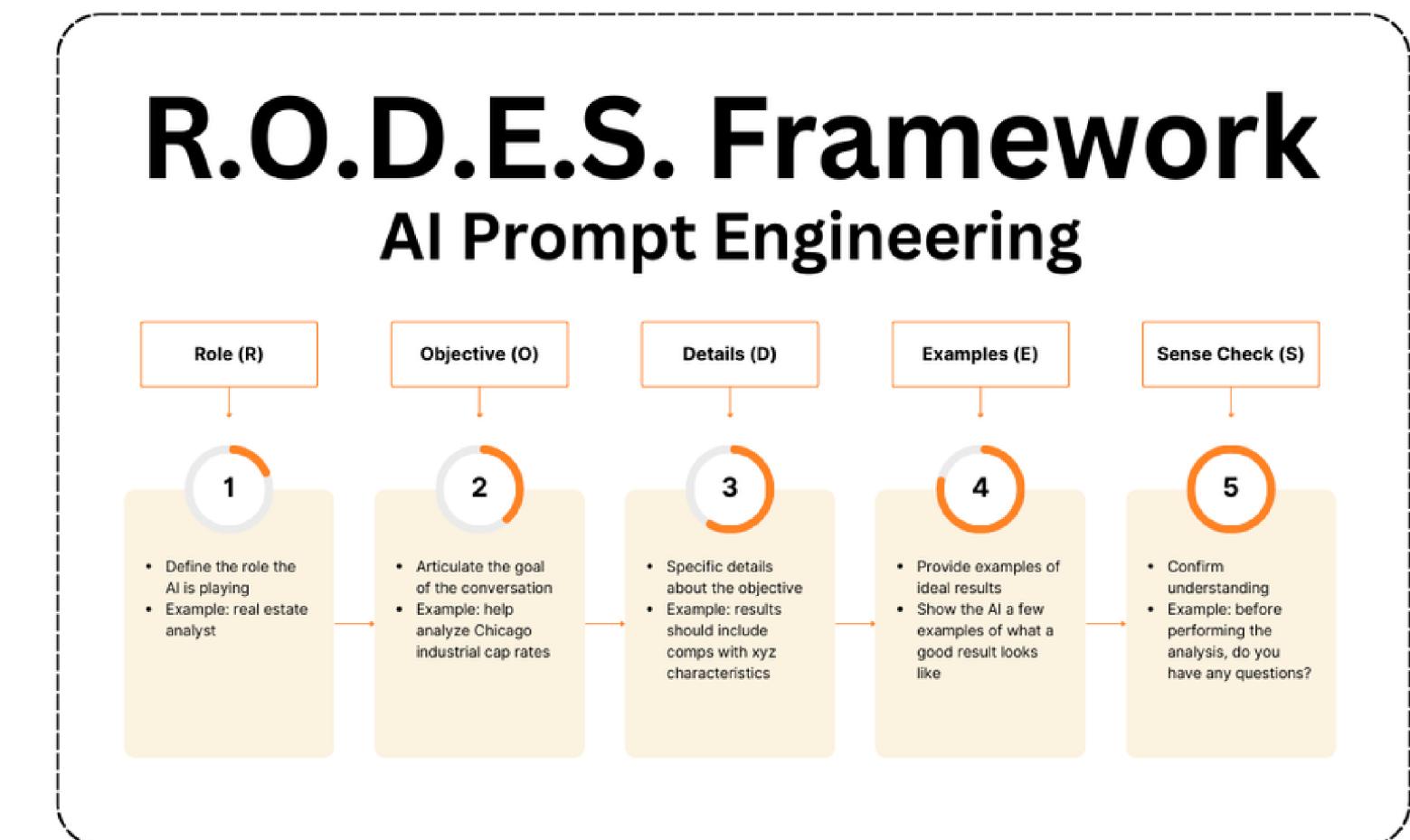
# Construyendo una solución con LLMs

## Desarrollo de instrucción (Prompt)

Cree una publicación en las redes sociales visualmente impactante y emocionalmente atractiva que refleje la experiencia de un especialista en marketing digital. El contenido debe ser altamente personalizado, fomentando la interacción a través de personas identificables.

Narraciones o preguntas que inviten a la reflexión.

La publicación debe ser adecuada para un público diverso, alineado con los valores de la marca, y optimizado para una alta participación en las plataformas de redes sociales.



# Construyendo una solución con LLMs

## Evaluación:

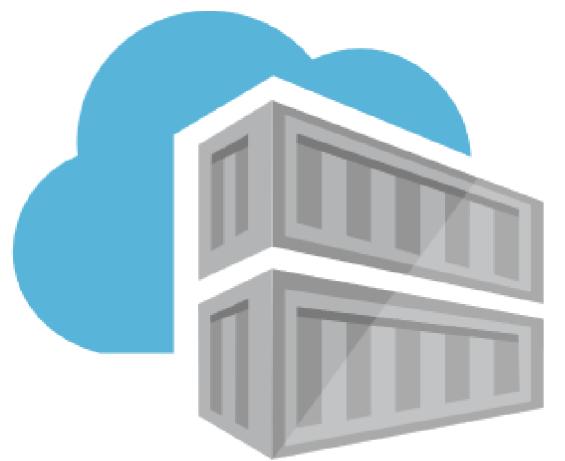
El modelo es capaz de responder basándose en la base de conocimiento creada.

Si quisiéramos que el modelo sea más “conversón” podemos ajustar parámetros dentro del modelo.

The screenshot shows a dark-themed web application titled "My Post Generator". At the top, it says "Welcome to 📧 My Post Generator ✨". Below that, a sub-headline reads "This is a simple tool to generate your post." A text input field is labeled "Enter the content for the post here 😊". Inside the input field, there is sample text about Streamlit: "A faster way to build and share data apps. Streamlit turns data scripts into shareable web apps in minutes. All in pure Python. No front-end experience required." A small red notification bubble with the number "1" is visible in the top right corner of the input field. Below the input field, a dropdown menu is set to "Captivating". A "Generate Post 🚀" button is present. The main content area displays the generated post: "Want to build and share your own data apps lightning-fast? Look no further than Streamlit! 💡". It continues with "With Streamlit, you can transform your data scripts into sleek and shareable web apps in just minutes, all using pure Python. Say goodbye to complex front-end development – no previous experience required! 🙌". Another section highlights "Whether you're a data enthusiast, analyst, or developer, Streamlit has got you covered. Share your insights, visualize your data, and create interactive experiences effortlessly. 🎨📊". Finally, it encourages users to "Ready to take your data projects to the next level? Give Streamlit a try and start building amazing apps today! ⚡💻". A footer at the bottom contains the hashtags "#Streamlit #DataApps #Python #Productivity".

# Construyendo una solución de GenAI

Despliegue:



Azure Container Registry



Azure Kubernetes Service(AKS)

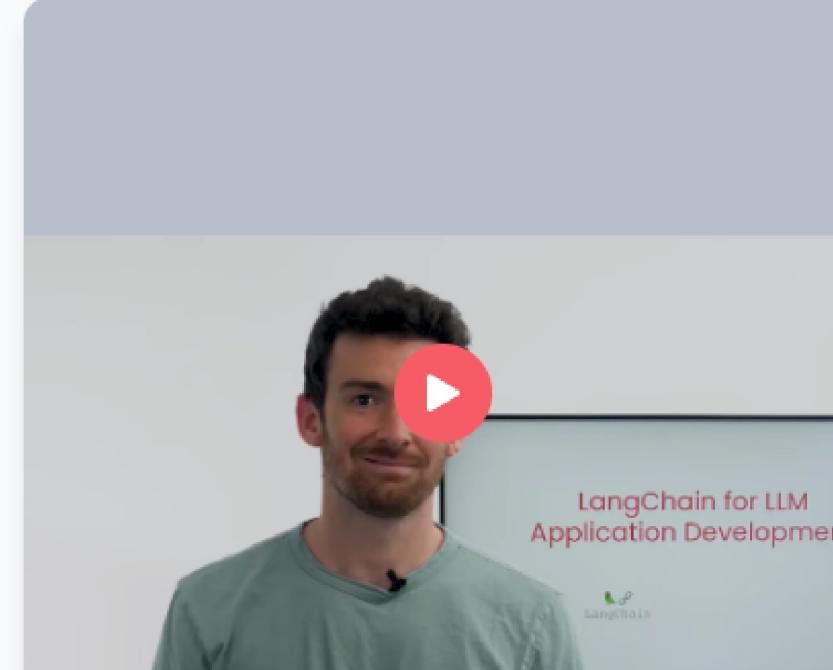


# Construyendo una solución de GenAI

## Monitoreo:

The screenshot shows the Azure portal's resources page. At the top, there's a navigation bar with icons for creating a resource, resource groups, Kubernetes services, Azure OpenAI, Subscriptions, Azure Databricks, Virtual machines, Azure Machine Learning, Marketplace, and More services. Below this is a section titled "Resources" with tabs for "Recent" (which is selected) and "Favorite". The "Recent" tab displays three items: a Resource group named "GENAL\_AMALIACONF\_2023" (last viewed "a few seconds ago"), a Kubernetes service named "aks-losy4gr5cu3ie" (last viewed "2 hours ago"), and a Subscription named "Pay-As-You-Go" (last viewed "2 hours ago"). There's also a "See all" link. At the bottom left is a "Navigate" button. On the right side of the screen, there's a large, semi-transparent monitoring chart. The chart has a timeline from "08 Nov" to "29 Nov" at the bottom. It features two stacked bars: a green bar representing "GPT-3.5 Turbo" and a magenta bar representing "Embedding models". A legend on the right side of the chart indicates that the green bar corresponds to "25 Nov" and "<\$0.01", while the magenta bar corresponds to "25 Nov" and "<\$0.01".

# Recursos para aprender más sobre LLMs



The thumbnail shows a man with a beard and short hair, wearing a green t-shirt, standing in front of a whiteboard. A red play button icon is overlaid on the top left of the video frame. The whiteboard behind him has the text "LangChain for LLM Application Development" and the LangChain logo.

**IN COLLABORATION WITH**  
LangChain 

## LangChain for LLM Application Development

The framework to take LLMs out of the box. Learn to use LangChain to call LLMs into new environments, and use memories, chains, and agents to take on new and complex tasks.

- Learn LangChain directly from the creator of the framework, Harrison Chase
- Apply LLMs to proprietary data to build personal assistants and specialized chatbots
- Use agents, chained calls, and memories to expand your use of LLMs

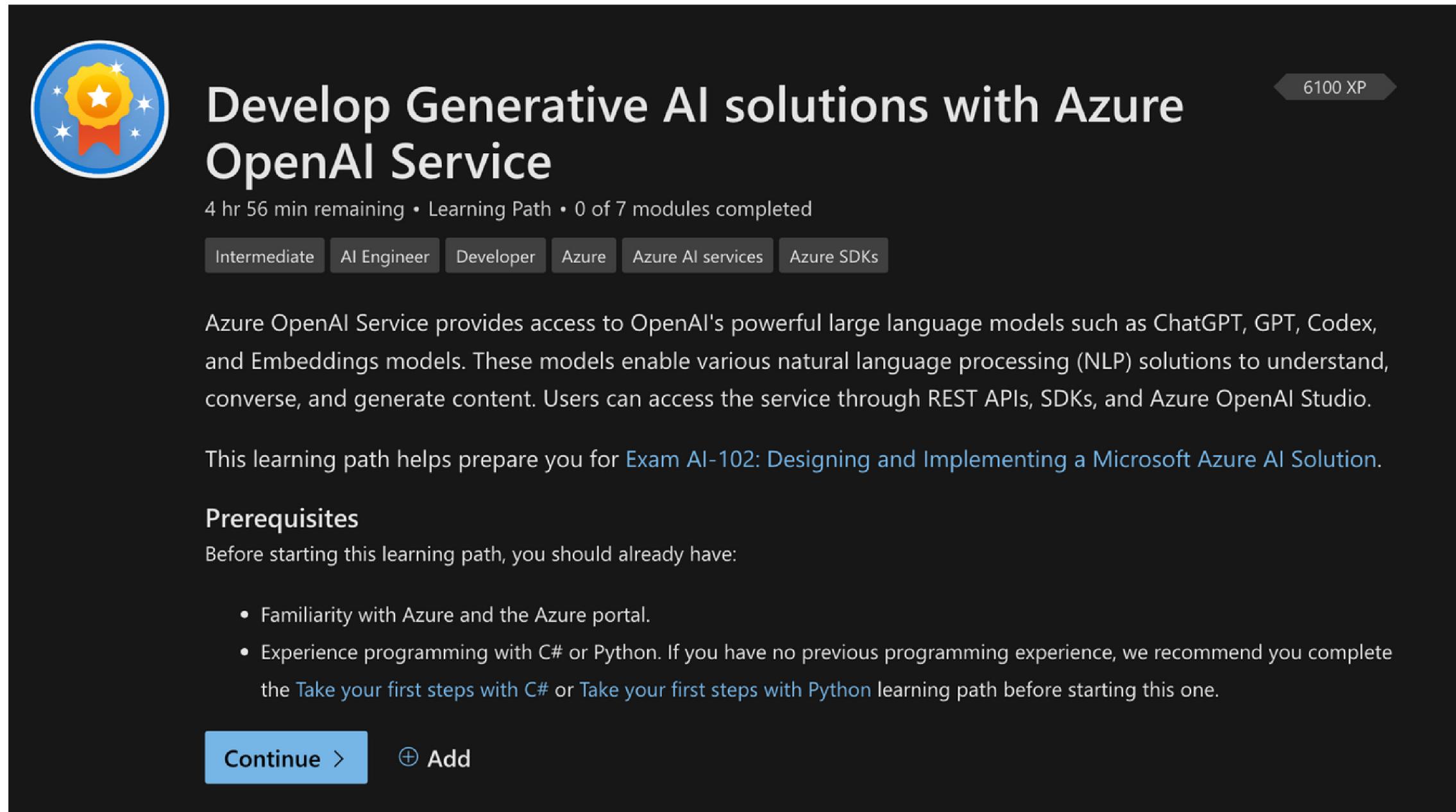
 Beginner  Harrison Chase, Andrew Ng

 Prerequisite recommendation: Basic Python

[Enroll For Free](#) [Learn more](#)

Enlace: [DeepLearningAI short courses](#)

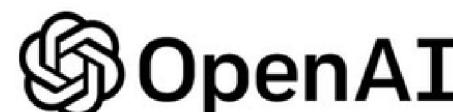
# Recursos para aprender más sobre LLMs



The screenshot shows a learning path titled "Develop Generative AI solutions with Azure OpenAI Service". It includes a circular icon with a gold ribbon and stars, a progress bar showing 6100 XP, and a timestamp of 4 hr 56 min remaining. The path has 0 of 7 modules completed. Below the title, there are tabs for Intermediate, AI Engineer, Developer, Azure, Azure AI services, and Azure SDKs. A description explains that Azure OpenAI Service provides access to OpenAI's powerful large language models like ChatGPT, GPT, Codex, and Embeddings. It mentions REST APIs, SDKs, and Azure OpenAI Studio. The path is designed to prepare for Exam AI-102: Designing and Implementing a Microsoft Azure AI Solution. Prerequisites include familiarity with Azure and the Azure portal, and experience with C# or Python. Buttons at the bottom allow users to "Continue >" or "+ Add".

Para obtener más información sobre cómo crear soluciones con el servicio Azure OpenAI.

Enlace: [Microsoft Learn](#)





# Gracias!

## Scanea el QR para acceder al app



## ¿Cómo contactarme?

- **LinkedIn:** Kevin Knights  
<https://www.linkedin.com/in/knightsk/>
- **Gmail:** kevin.k.knights@gmail.com