

Building a Cutting-Edge AI Lab

24.10.2024

Machine Learning vs Generative AI

Traditional ML: creating and training models for specific purposes

Resource-intensive training and tuning required.

Generative AI + Retrieval-Augmented Generation (RAG)

Pre-trained generative models retrieve information from documents and databases. No training required.

Large Language Models vs Small Language Models

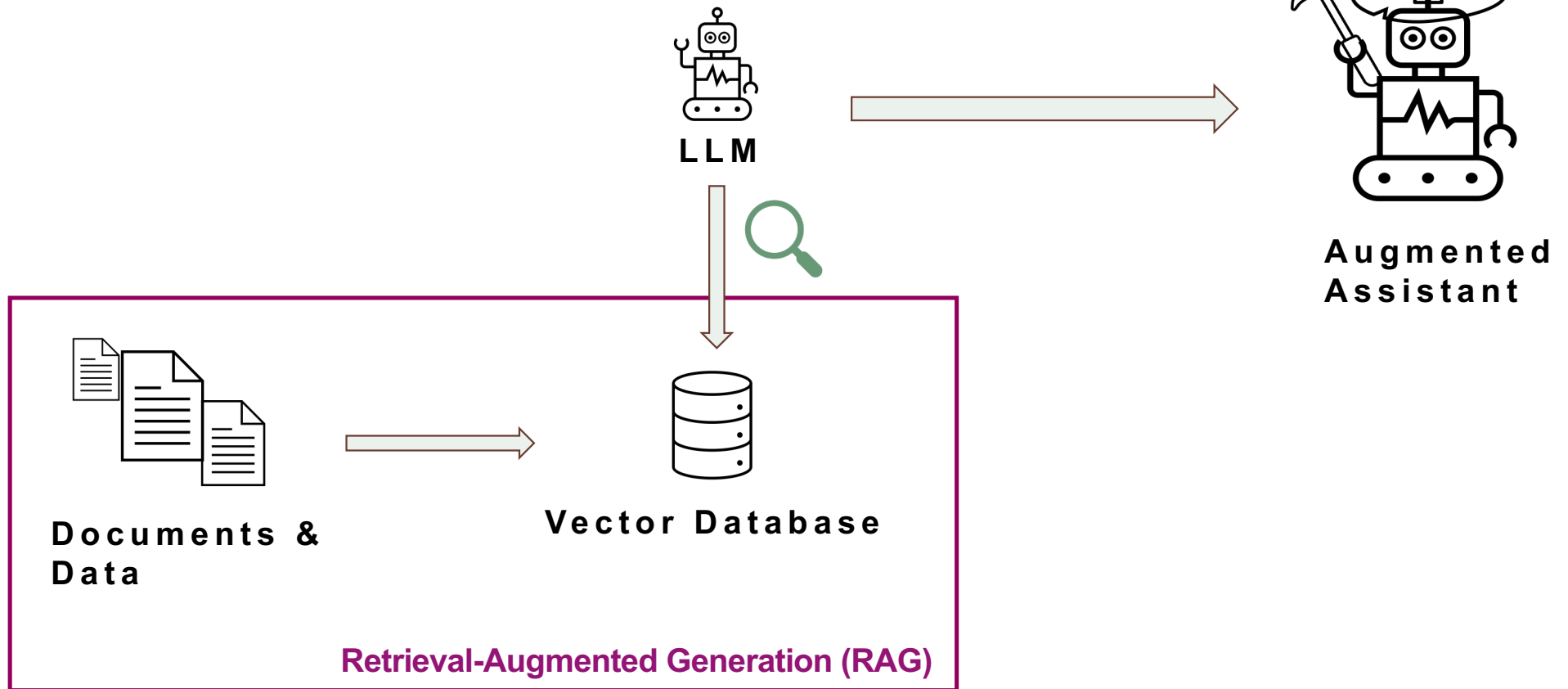
Large Language Models

LLMs like **GPT-4** are massive neural networks trained on vast amounts of text data, enabling them to generate human-like responses on a wide range of subjects. They excel at **open-ended conversation, summarization, and content generation**.

Small Language Models

Smaller, more specialized language models like **phi-3** have a narrower focus, often **trained on specific domains or tasks**. They tend to be faster, more efficient, and require less computational power, making them suitable for edge devices or resource-constrained environments.

AI Assistants



Use Cases

User Facing Assistants



Customer Support Assistant



**Personalized Healthcare Assistant
for Doctors, based on patient
history**



**Educational Assistant for
Professors**

Technical Assistants



**Medical records and
diagnosis to XML/JSON**



**Product catalog details
structuring**



Email data extraction

Building Your GenAI Lab: your own hardware

Infrastructure

Processor:

- Intel Core i7 (11th or 12th generation)
- AMD Ryzen 7 (5800H or 6800H)

GPU: NVIDIA RTX 3060 (6GB VRAM)

RAM: 32 GB

Storage: 1 TB SSD

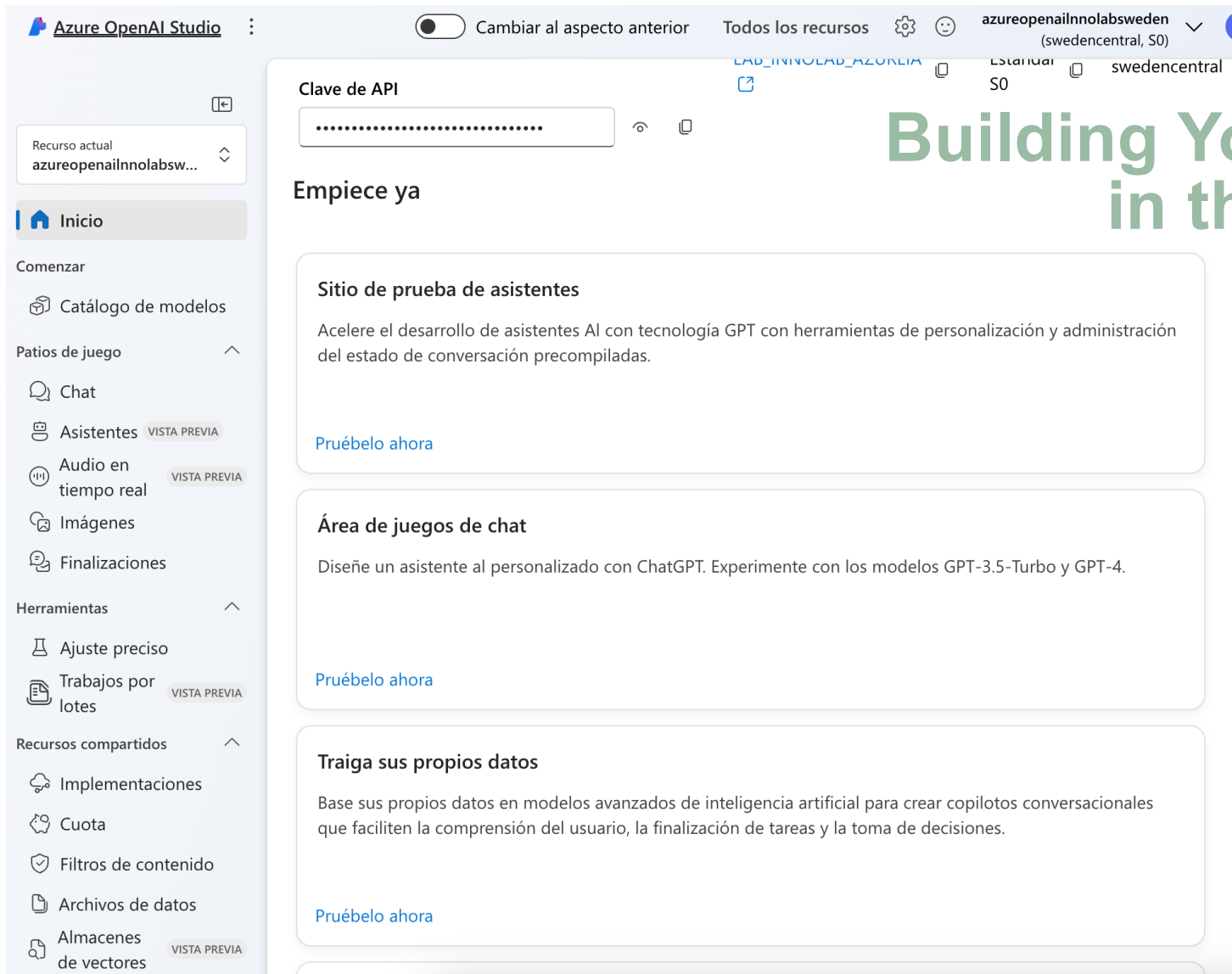
www.semicstore.es

LLM Models



Hugging Face

<https://huggingface.co/models>



Building Your GenAI Lab: in the cloud

Azure OpenAI Studio

Área de juegos de asistentes

Cambiar al aspecto anterior

Todos los recursos

azureopenailnnolabswe...
(swedencentral, S0)

Recurso actual
azureopenailnnolabsw...

Inicio

Comenzar

Catálogo de modelos

Patios de juego

Chat

Asistentes

VISTA PREVIA

Audio en tiempo real

VISTA PREVIA

Imágenes

Finalizaciones

Herramientas

Ajuste preciso

Trabajos por lotes

VISTA PREVIA

Recursos compartidos

Implementaciones

Cuota

Filtros de contenido

Archivos de datos

Almacenes de vectores

VISTA PREVIA

Área de juegos de asistentes

+ Nuevo asistente

↑ Seleccionar asistente

Eliminar

</> Ver código

Configuración

Ocultar

Implementación *

+ Crear nueva implementación

gpt-4o-mini-SAIN (version:2024-07-18)

Nombre del asistente ⓘ

atlas-configurator-assistant

Id. del asistente: asst_ayPS9EtZZJV5AfQf7IYzFrwM

Solicitud

Herramientas

Parámetros

Instrucciones ⓘ

You are helping a user to choose a PC configuration according the following questions:

1. What do you need the PC for? In which field you work in (sales, IT, programming, etc)?

2. Do you need any accessories like a keyboard or mouse?

3. Do you need a warranty for the product?

Borrar chat

Registros

Archivos de subproceso

Respuesta JSON

580 tokens ⓘ

Nuevo subproceso iniciado

thread_FZYuROjtq8Bfppwqj251VzGt ⓘ

Hello I need a PC

Sure! I have a few questions to help you choose the right PC configuration:

1. **What do you need the PC for?** In which field do you work (sales, IT, programming, etc.)?

2. **Do you need any accessories** like a keyboard or mouse?

3. **Do you need a warranty** for the product?

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

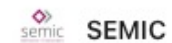
Los mensajes en el área de juegos de Assistants son visibles para cualquier persona con acceso a este recurso y mediante la API.

+

Azure AI Studio



Bianca Pădurean ✓
CIO / COO



www.linkedin.com/in/bianca-padurean

bpadurean@semic.es

GRACIAS



www.semic.es | info@semic.es

MADRID	BARCELONA	LLEIDA	GIRONA	MANRESA	SEVILLA	VALENCIA	ZARAGOZA	CANARIAS	ANDORRA
T. 91 789 65 90	T. 93 492 41 50	T. 973 28 08 00	T. 93 492 41 50	T. 93 492 41 50	T. 955 25 04 70	T. 91 789 65 90	T. 91 789 65 90	T. 91 789 65 90	T. +376 80 82 90

