

# Uso de modelos de lenguaje de gran tamaño (LLM) de código abierto

*¿Es posible hacerlo "On Premise"?*

Isaac Vidal Daza

Apoyo a la Docencia  
Centro de Servicios de Informática y Redes de Comunicaciones  
Universidad de Granada

28-05-2024



UNIVERSIDAD  
DE GRANADA

# Teaching Support Department

## Members

- Francisco Romera Juárez. (Head)
- Fernando López Álvarez.
- José Guerrero Peregrina.
- Antonio Cano Ruano.
- Rodrigo González Gálvez.
- Domingo Baca Ruiz.
- Leire Melchor López.
- Isaac Vidal Daza.



UNIVERSIDAD  
DE GRANADA

# Teaching Support Department

## Services

- Computers Classrooms Management.
- Software Deployment.
- Virtual Desktop Infrastructure.
- Microsoft 365 Management.
- Systems and Services Deployment.

## Inventory

- 25 Faculties (Ceuta y Melilla).
- 128 Computers Classrooms.
- 3418 PC's.
- 200 Virtual Desktops.

## Users

- 53000 Students.
- 3600 Teaching Staff.
- 2200 Administration Staff.



UNIVERSIDAD  
DE GRANADA

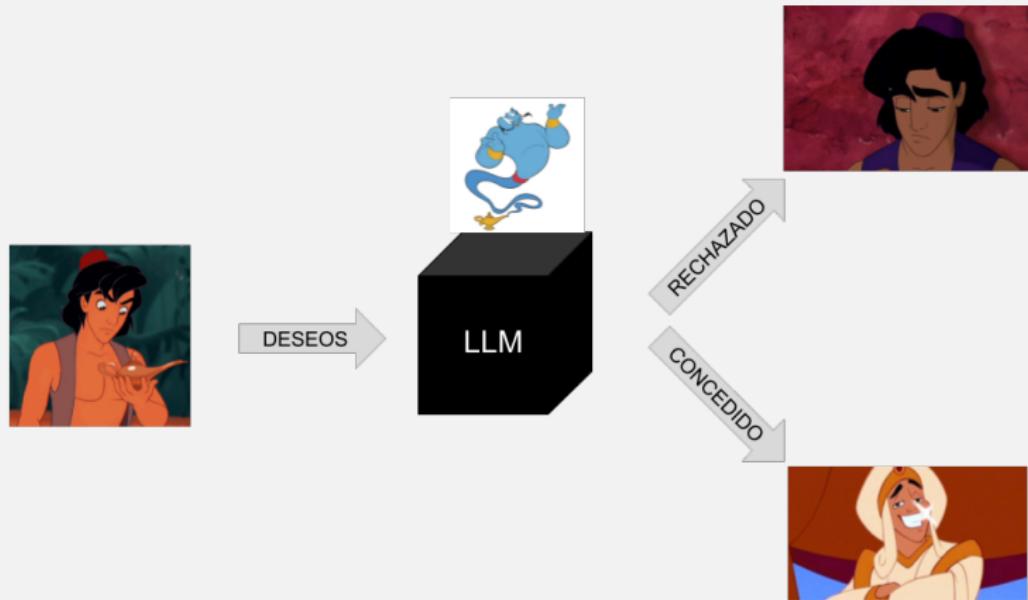
# Inteligencia Artificial (según los "Mass Media")



UNIVERSIDAD  
DE GRANADA

# Modelos de Lenguaje

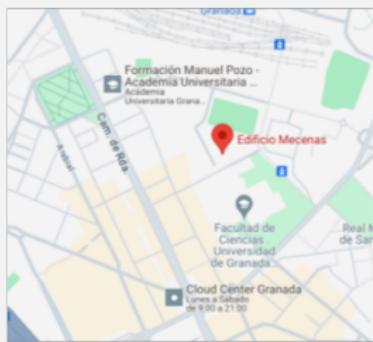
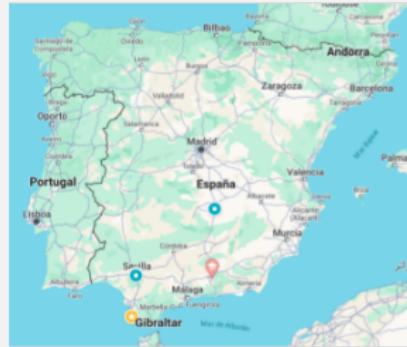
## Lámpara Mágica



UNIVERSIDAD  
DE GRANADA

# Modelos de Lenguaje

## Instrucciones Precisas



UNIVERSIDAD  
DE GRANADA

# ¿Cómo pasar de Espectador a Ilusionista?

## Utilizando modelos LLM Open Source

- Llama3\*
- Mistral

## ¿Dónde se encuentran?

- <https://huggingface.co/>
- <https://mistral.ai/>

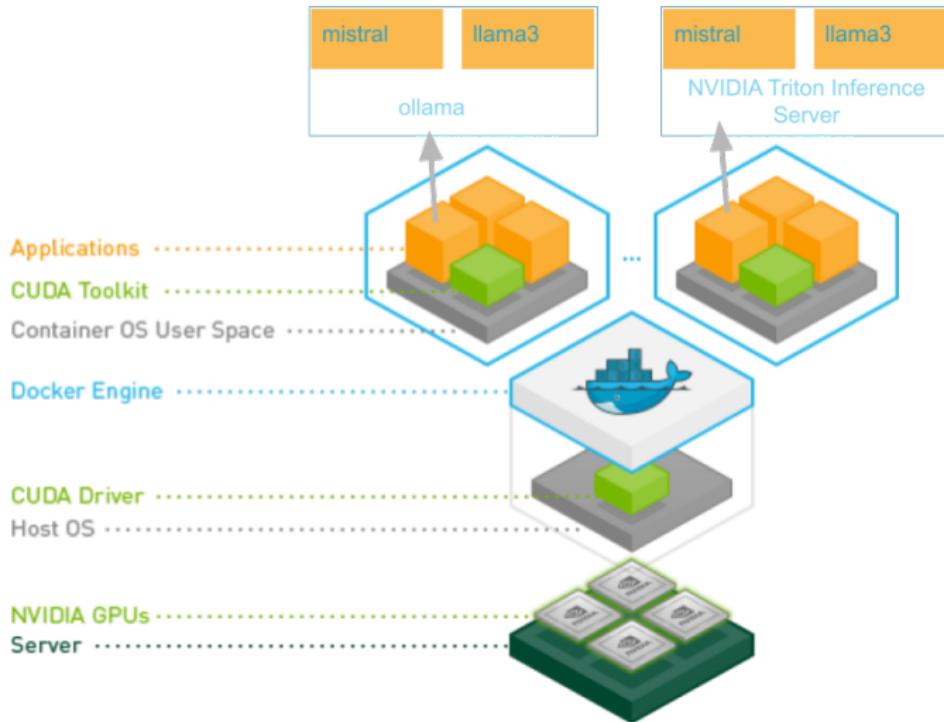
## ¿Dónde ejecutan?

- On Premise
- Nube



UNIVERSIDAD  
DE GRANADA

# ¿Cómo ejecutarlos? CPUs y GPUs



# Interfaz Gráfica (unas 40 líneas de código)

Listing 1: Código usando el framework Chainlit

```
1      @cl.on_chat_start
2      async def on_chat_start():
3          model = Ollama(base_url="http://localhost:11434", model="llama3")
4          prompt = ChatPromptTemplate.from_messages(
5              [
6                  (
7                      "system",
8                          "You are a helpful assistant. You must always respond to
9                             Spanish questions if you receive a question in any other
10                            language. "
11                  ),
12                  ("human", "{question}"),
13              ]
14          )
15          runnable = prompt | model | StrOutputParser()
16          cl.user_session.set("runnable", runnable)
```



# Interfaz Gráfica (unas 40 líneas de código)

Listing 2: Código usando el framework Chainlit

```
1      @cl.on_message
2      async def on_message(message: cl.Message):
3          runnable = cl.user_session.get("runnable") # type: Runnable
4
5          msg = cl.Message(content="")
6
7          async for chunk in runnable.astream(
8              {"question": message.content},
9              config=RunnableConfig(callbacks=[cl.LangchainCallbackHandler()]),
10         ):
11             await msg.stream_token(chunk)
12
13         await msg.send()
```



# Resultado 40 líneas de código

## Asistente Personalizado

The screenshot shows a Mozilla Firefox browser window with a single tab titled "Chatbot — Navegación privada de Mozilla Firefox". The address bar displays "localhost:8000". The main content area features a header with the CSIRC logo and navigation links for "Chat" and "Readme". A large heading says "Bienvenido al Asistente IA del CSIRC". Below it, a message reads "Hola, soy MANGUS ¿en qué puedo ayudarte?". At the bottom, there is a message input field with placeholder text "Type your message here..." and a send button.



UNIVERSIDAD  
DE GRANADA

# Resultado 40 líneas de código

## Asistente Personalizado

Chatbot — Navegación privada de Mozilla Firefox

Chatbot Navegación privada

localhost:8000

CSIRC Chat Readme New Chat

You 18:23 Escribe una aplicación en python que sirva para cambiar una contraseña en LDAP Took 1 step

Chatbot 18:23 ¡Claro! Para crear una aplicación en Python que cambie la contraseña en LDAP, podemos utilizar la biblioteca ldap3 y la función modify para actualizar el atributo userPassword . Aquí te dejo un ejemplo básico:

```
python
import ldap3

# Configuración de la conexión a LDAP
server = 'ldap://localhost' # Cambia por tu servidor LDAP
basedn = 'dc=example,dc=com' # Cambia por tu base de datos LDAP
username = 'cn=admin,dc=example,dc=com' # Cambia por tu usuario administrador
password = 'adminpass' # Cambia por tu contraseña administradora
```

Type your message here...

Built with Chainlit



UNIVERSIDAD  
DE GRANADA

# Resultado 40 líneas de código

## Asistente Personalizado

Chatbot — Navegación privada de Mozilla Firefox

Chatbot    Navegación privada

localhost:8000

CSIRC Chat Readme New Chat

```
password = 'adminpass' # Cambia por tu contraseña administradora

# Conectamos a la base de datos LDAP
conn = ldap3.Connection(server, user=username, password=password, authentication='regular')

# Damos los detalles del usuario y la nueva contraseña
username_old = 'cn=user1,dc=example,dc=com'
new_password = 'nuevacontraseña123'

try:
    # Buscamos el usuario en LDAP
    entry = conn.search(search_scope='one', base=basedn, filter='(&(objectClass=person)(cn={}))'.format(username))

    # Actualizamos la contraseña del usuario
    conn.modify(entry[0].dn, {'userPassword': [(ldap3.MOD_REPLACE, new_password.encode('utf-8'))]})

    print('Contraseña actualizada con éxito!')
except Exception as e:
    print('Error:', str(e))
finally:
    conn.unbind()

Ten en cuenta que debes reemplazar los valores de server, basedn, username y password por tus
```

Type your message here...

Built with Chainlit



UNIVERSIDAD  
DE GRANADA

# ¿Puede tener memoria nuestro Asistente IA?

## ¿Base de Datos SQL?

The screenshot shows a video player interface. At the top, there is a large image of a black and white cat sitting on a wooden surface against a green wall. Three arrows point from this main image down to three smaller images below it: a ginger cat walking on a yellow background, a light-colored cat sitting on a teal background, and a close-up of a brown tabby cat's face. To the right of these images is a table representing the data in a SQL database:

animal	color	tags
cat	black, white	cute green eyes green background

At the bottom of the video player, there is a progress bar showing "1:08 / 4:22" and the text "Why do we need vector databases". The video player has standard controls for play/pause, volume, and navigation.

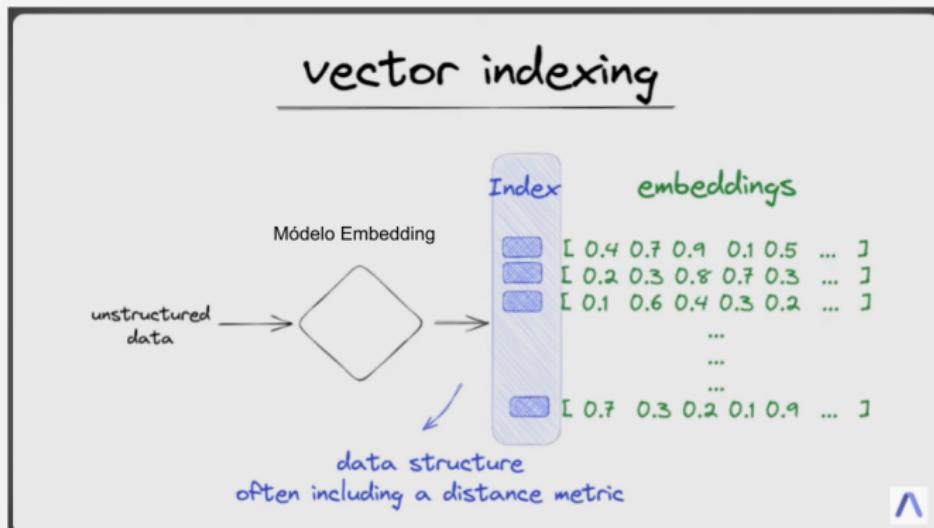
<https://www.youtube.com/watch?v=dN0lsF2cvm4>



UNIVERSIDAD  
DE GRANADA

# ¿Puede tener memoria nuestro Asistente IA?

## Bases de Datos Vectoriales



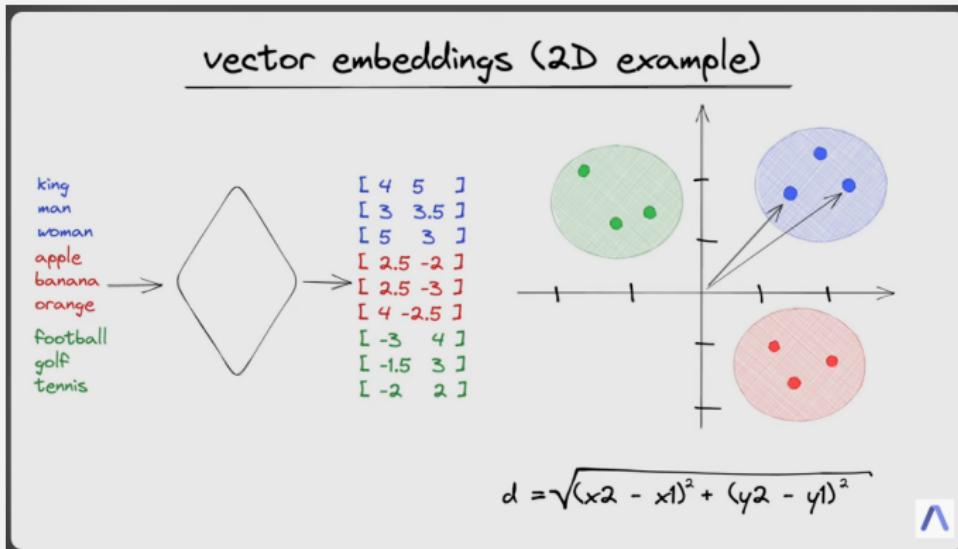
<https://www.youtube.com/watch?v=dN0lsF2cvm4>



UNIVERSIDAD  
DE GRANADA

# ¿Puede tener memoria nuestro Asistente IA?

## Bases de Datos Vectoriales



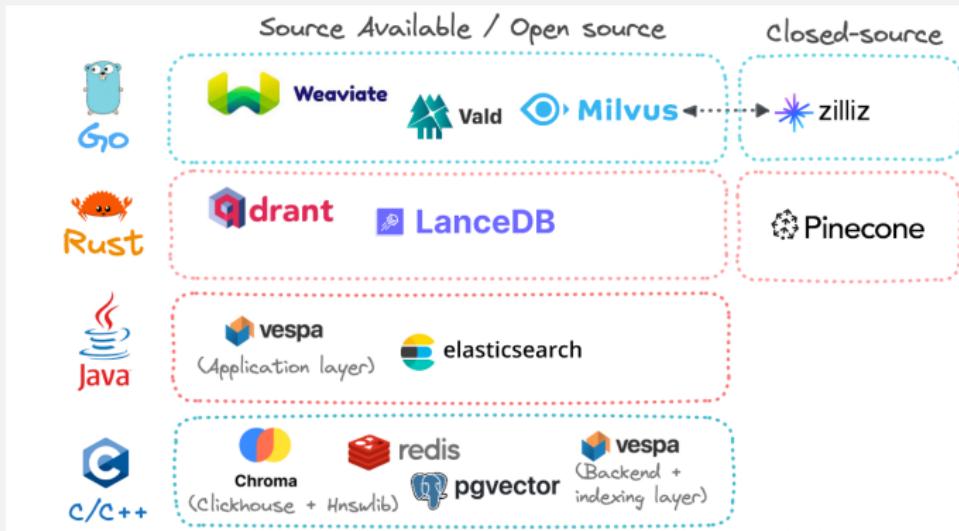
<https://www.youtube.com/watch?v=dN0lsF2cvm4>



UNIVERSIDAD  
DE GRANADA

# ¿Puede tener memoria nuestro Asistente IA?

## Bases de Datos Vectoriales



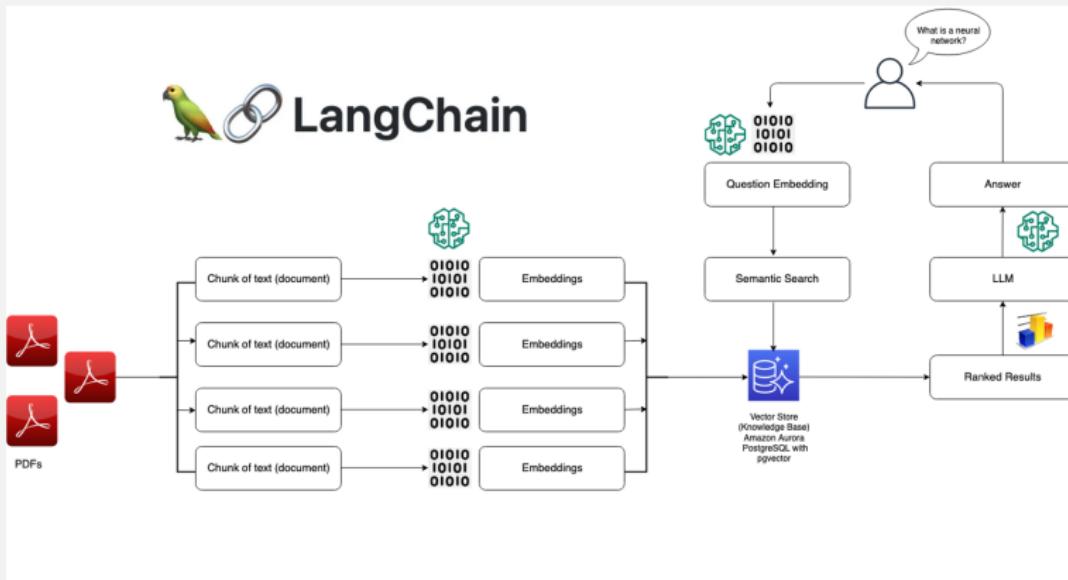
<https://thedataquarry.com/posts/vector-db-1/>



UNIVERSIDAD  
DE GRANADA

# LLMs y BD Vectoriales: Retrieval-Augmented Generation

## Frameworks LLM: Langchain y/o LlamaIndex



<https://www.langchain.com/langchain>



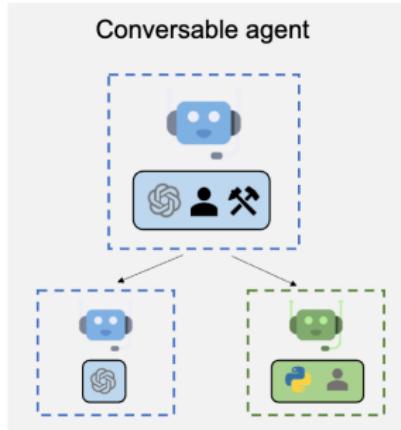
UNIVERSIDAD  
DE GRANADA

# Agentes

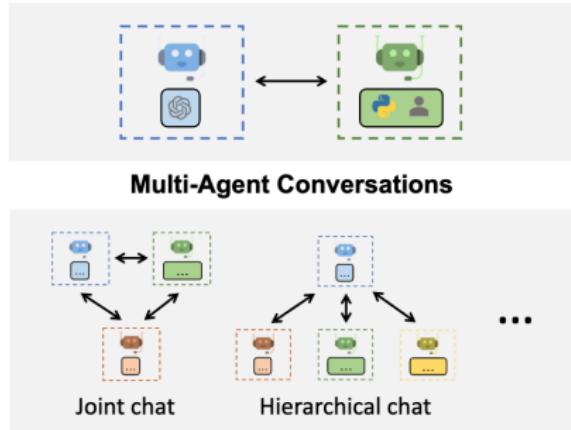


UNIVERSIDAD  
DE GRANADA

# Framework Autogen



**Agent Customization**



**Flexible Conversation Patterns**

<sup>0</sup><https://github.com/microsoft/autogen>



UNIVERSIDAD  
DE GRANADA

# Preguntas

## Presentación y Código

<https://github.com/isvida/2024-RedIRIS>



UNIVERSIDAD  
DE GRANADA