

LLM COMO SERVICIO

MiIA UGR (Mi Inteligencia Artificial)

Isaac Vidal Daza

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

22-05-2025



UNIVERSIDAD
DE GRANADA

Jt
RedIRIS

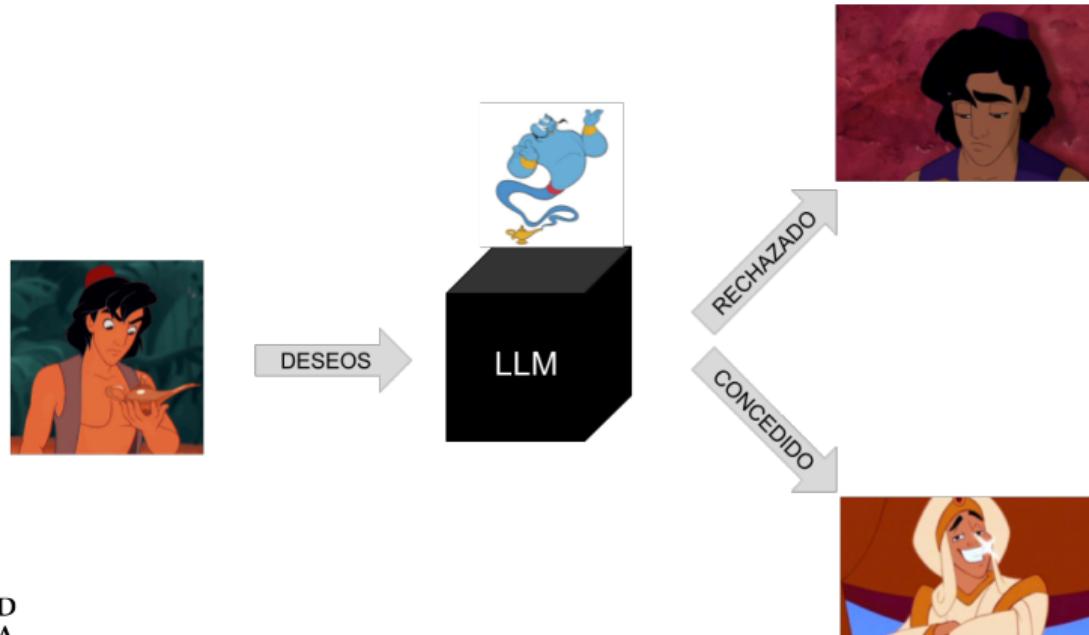
Inteligencia Artificial (según los "Mass Media")



UNIVERSIDAD
DE GRANADA

Modelos de Lenguaje

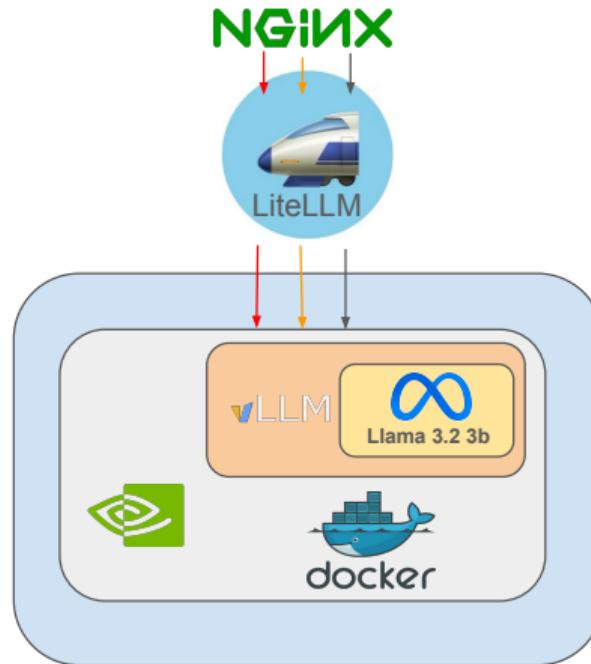
Lámpara Mágica



UNIVERSIDAD
DE GRANADA

Core MilA

Componentes del Servicio



UNIVERSIDAD
DE GRANADA

LiteLLM: proxy LLM

Definición de API keys: Virtual Keys

The screenshot shows the LiteLLM web application interface. At the top, there's a navigation bar with the LiteLLM logo, a 'Get enterprise license' button, and an 'Admin' button. On the left, a sidebar menu includes 'Virtual Keys' (which is highlighted in blue), 'Test Key', 'Models', 'Usage', 'Teams', 'Internal Users', 'Logging & Alerts', 'Caching', 'Budgets', 'Router Settings', 'Pass-Through', 'Admin Settings', 'API Reference', and 'Model Hub'. The main content area is titled 'CSIRC' and displays 'Team ID: 593c1594-9c94-4268-888e-ba80bec4072c'. It shows 'Total Spend \$0.0000' and 'Max Budget No limit'. Below this, a table lists three API keys:

Key Alias	Secret Key	Created	Expires	Spend (USD)	Budget (USD)	Budget Reset	Models	Rate Limits
...@ugr...	sk-...J9EA	10/12/2024	Never	0.0000	Unlimited	Never	All Team Models	TPM: Unlimited RPM: Unlimited
...@ugr...	sk-...JzJA	10/12/2024	Never	0.0000	Unlimited	Never	All Team Models	TPM: Unlimited RPM: Unlimited
...@ugr...	sk-...w97w	10/12/2024	Never	0.0000	Unlimited	Never	All Team Models	TPM: Unlimited RPM: Unlimited

At the bottom of the table area is a blue button labeled '+ Create New Key'. Below the table, there's a section titled 'Select Team' with the note: 'If you belong to multiple teams, this setting controls which team is used by default when creating new API Keys.' A dropdown menu is shown with 'CSIRC' selected. There's also a note about the 'Default Team': 'If no team_id is set for a key, it will be grouped under here.'



UNIVERSIDAD
DE GRANADA

LiteLLM: proxy LLM

Definición de modelos y control de accesos.

The screenshot shows the LiteLLM web application interface. At the top, there is a navigation bar with links for 'Virtual Keys', 'All Models' (which is highlighted in blue), 'Add Model', '/health', 'Models', 'Model Analytics', and 'Model Retry Settings'. On the right side of the header, there are buttons for 'Get enterprise license' and 'Admin'. Below the header, a message 'Last Refreshed: 20/5/2025, 20:30:29' is displayed. The main content area is titled 'Models' and contains a table with three rows of data. The columns are: 'Public Model Name', 'Provider', 'LiteLLM Model', 'API Base', 'Input Price /1M Tokens (\$)', 'Output Price /1M Tokens (\$)', 'Created At', 'Created By', and 'Status'. The first row has values: 'mangus-code2', 'hosted_vilm', 'hosted_vlm/Qwen/Qwe...', 'http://orca1.u gr.es:', '-', '-', 'DB Model', and icons for edit and delete. The second row has values: 'mangus-code', 'hosted_vilm', 'hosted_vlm/Qwen/Qwe...', 'http://orca1.u gr.es:', '-', '-', 'DB Model', and icons for edit and delete. The third row has values: 'mangus', 'hosted_vilm', 'hosted_vlm/unslot h/...', 'http://orca1.u gr.es:', '-', '-', 'DB Model', and icons for edit and delete. To the left of the main content area, there is a sidebar with links for 'Test Key', 'Models' (which is highlighted in blue), 'Usage', 'Teams', 'Internal Users', 'Logging & Alerts', 'Caching', 'Budgets', 'Router Settings', 'Pass-Through', 'Admin Settings', 'API Reference', and 'Model Hub'.

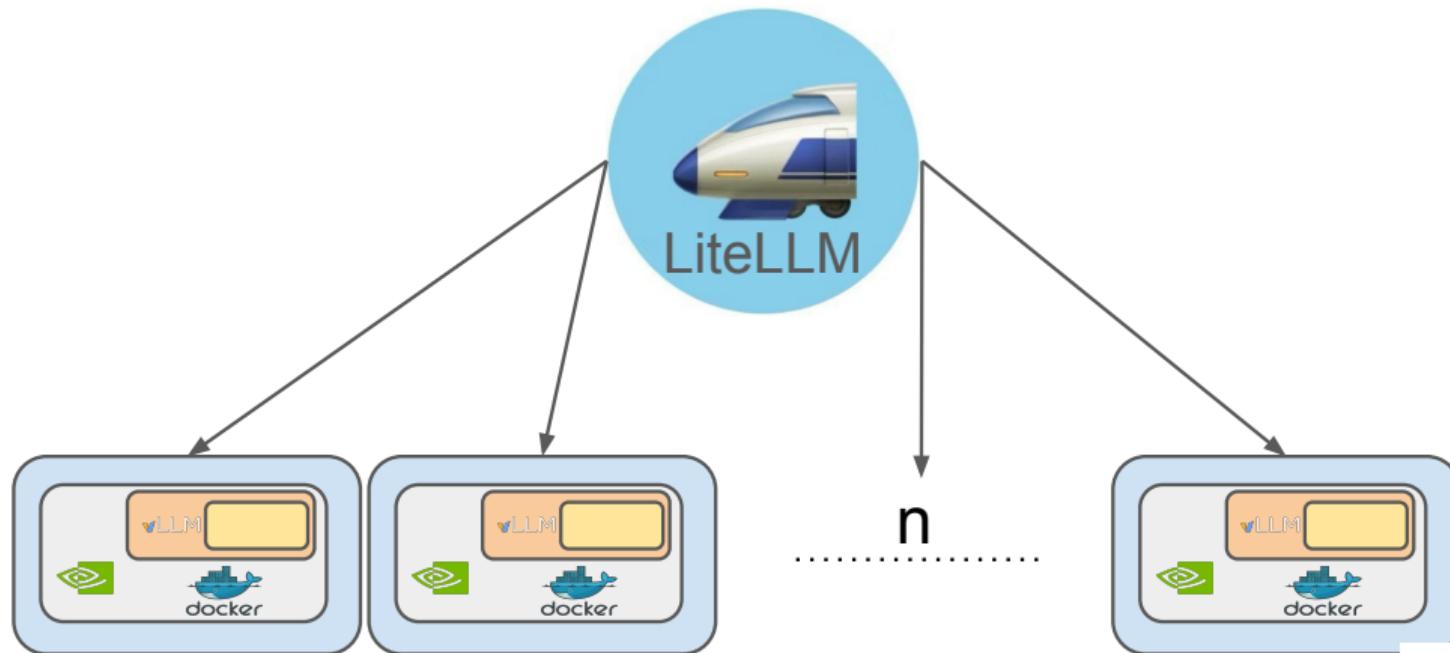
Public Model Name	Provider	LiteLLM Model	API Base	Input Price /1M Tokens (\$)	Output Price /1M Tokens (\$)	Created At	Created By	Status
mangus-code2	hosted_vilm	hosted_vlm/Qwen/Qwe...	http://orca1.u gr.es:	-	-			DB Model
mangus-code	hosted_vilm	hosted_vlm/Qwen/Qwe...	http://orca1.u gr.es:	-	-			DB Model
mangus	hosted_vilm	hosted_vlm/unslot h/...	http://orca1.u gr.es:	-	-			DB Model



UNIVERSIDAD
DE GRANADA

LiteLLM: proxy LLM

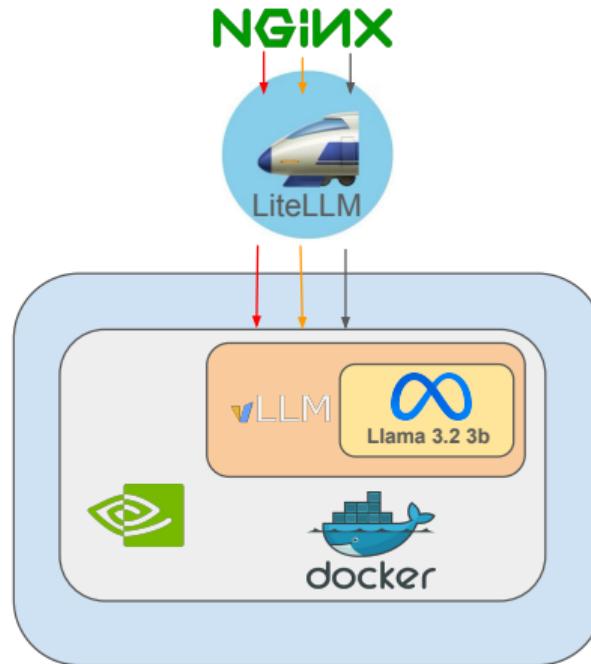
Escalado del servicio LLM

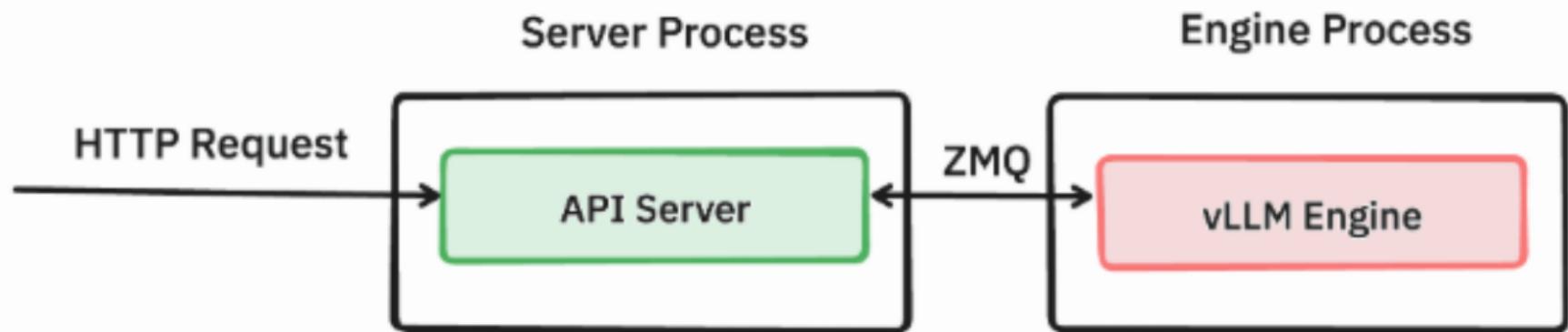


UNIVERSIDAD
DE GRANADA

Core MilA

Componentes del Servicio





Servicios Ofrecidos con MilA Core

MiA Chat (Open WebUI)

The screenshot shows the MiA Chat (Open WebUI) interface. On the left is a sidebar with the following items:

- Nuevo Chat
- Notas
- Espacio de Trabajo
- Buscar
- Chats (with a dropdown arrow)
- CHEMISTRY
- Hoy
- Título de Ingeniero 🎓
- ***Visible Text Transcription:**\n
- hola
- Nuevo Chat
- PhD Research Data Analysis
- LaTeX Conversion and Explan
- 7 días previos

Below the sidebar is a user profile section:

ID Isaac Vidal Daza

The main area displays a message from the system:

mangus Establecer como Predeterminado

INFO Bienvenido al asistente virtual del Servicio de Apoyo a la Docencia (CSIRC) de la Universidad de Granada. Este servicio es solo para usos docentes, dirigido exclusivamente a estudiantes y profesores.

Below this is a large message from the AI model:

mangus
Modelo LLM basado en Llama 3.2 3B Instruct

¿Cómo puedo ayudarte hoy?

+ Búsqueda Web Interprete de Código

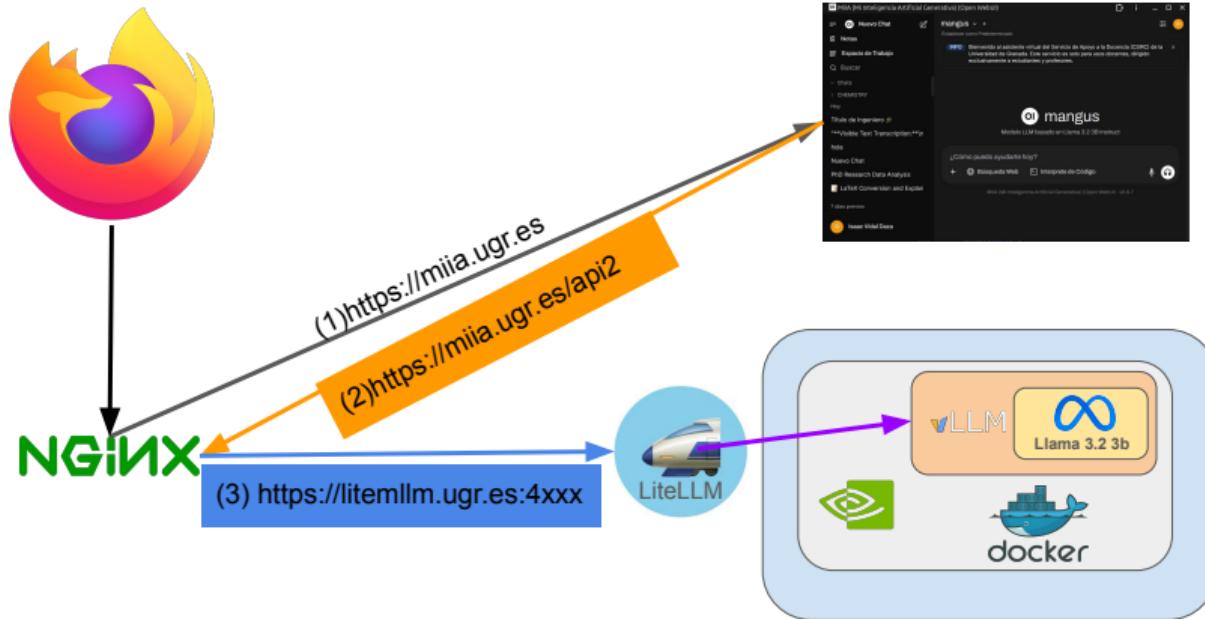
MiA (Mi Inteligencia Artificial Generativa) (Open WebUI) · v0.6.7



UNIVERSIDAD
DE GRANADA

Servicios Ofrecidos con MilA Core

MilA Chat (*Open WebUI*): Arquitectura



UNIVERSIDAD
DE GRANADA

Servicios Ofrecidos con MilA Core

ThunderAI: Extensión de Thunderbird para correo electrónico

The screenshot shows the Mozilla Thunderbird interface with the ThunderAI extension. The window title is "Escribir: (sin asunto) - Thunderbird". The menu bar includes Archivo, Editar, Ver, Insertar, Formato, Opciones, Herramientas, Ayuda, Envíar, Cifrar, Ortografía, Guardar, Contactos, Adjuntar, and AI. The main window shows a message being composed to "isvda@ugr.es". The message body contains the text "Hola, No tengo acceso al correo electrónico, ¿puedes ayudarme? Saludos.". A tooltip from the AI extension says, "Attempting to connect to the OpenAI Compatible API Local Server using the host "https://mila.ugr.es/api2" and model "mangus"..." Below the message body, there is an "Information" section with the instruction: "Rewrite the following text to be more polite. Reply with only the re-written text and with no extra comments or other text. Reply in the same language. "Hola, No tengo acceso al correo electrónico, ¿puedes ayudarme? Saludos."". A suggested rewrite is shown: "Hola, ¿podrías ayudarme con mi correo electrónico? Muchas gracias." There are buttons for "Use this answer", "Show differences", and "Close". Below this, a "Differences between the original and the modified text" section shows the changes: "Hola, No tengo acceso al correo electrónico, ¿puedes ayudarme con mi correo electrónico? SaludosMuchas gracias." The ThunderAI logo is visible in the bottom right corner of the message window.



UNIVERSIDAD
DE GRANADA

Servicios Ofrecidos con MiLA Core

Continue: Extensión de Visual Studio Code para desarrollo asistido por IA

The screenshot shows the LaTeX Workshop extension in Visual Studio Code. The main editor pane displays a LaTeX document with several frames. The current cursor is at line 214, which starts a frame titled "Atención de los Transformadores: un enfoque matemático". The extension's interface includes a sidebar with "CONTINUE" and "CHAT" tabs, and a floating panel with the message: "añade una diapositiva utilizando latex beamer que explique el mecanismo de atención de los transformadores. No incluyas imágenes". The bottom status bar shows compiler logs and error messages.

```
\begin{frame}[Servicios Ofrecidos con MiLA Core]
    \Includegraphics[height=6cm]{Imagenes/thunderAI.png}
\end{center}
\end{frame}

\begin{frame}[Servicios Ofrecidos con MiLA Core]
    \framesubtitle[\textbf{Continue}]: Extensión de Visual Studio Code para desarrollo asistido por IA

    \begin{center}
        \Includegraphics[height=6cm]{imagenes/vscode_continue2.png}
    \end{center}
\end{frame}

\begin{frame}[Atención de los Transformadores: un enfoque matemático]
    \begin{center}
        La atención de los transformadores se basa en la idea de que cada token de entrada es un vector de dimensión \$d\$.
        El objetivo es transformar estos vectores en una salida de dimensión \$D\$ que representa la probabilidad de que el token de salida sea un determinado símbolo.
        La atención se logra mediante una capa de transposición que permite a los tokens de entrada interactuar entre sí, y una capa de dot producto que combina estos tokens para producir la salida.
        En resumen, la atención de los transformadores se puede resumir en la siguiente ecuación:
    \end{center}

```



UNIVERSIDAD
DE GRANADA

Atención de los Transformadores: un enfoque matemático

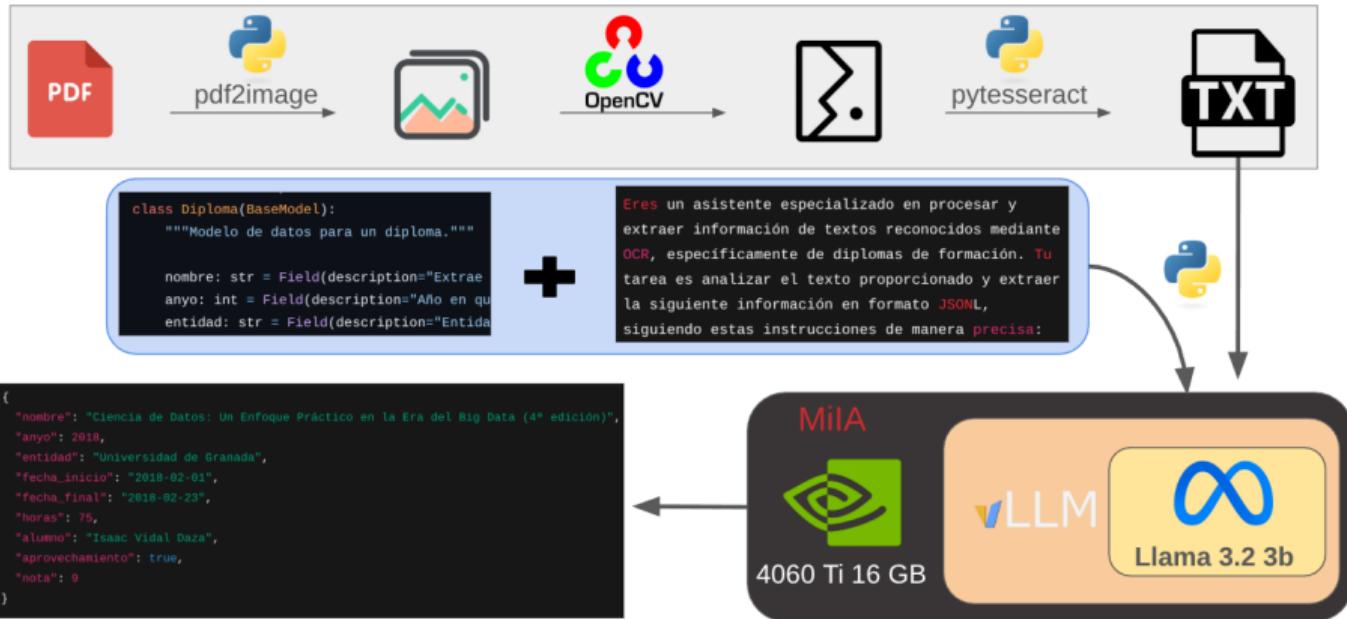
La atención de los transformadores se basa en la idea de que cada token de entrada es un vector de dimensión d . El objetivo es transformar estos vectores en una salida de dimensión D que representa la probabilidad de que el token de salida sea un determinado símbolo. La atención se logra mediante una capa de transposición que permite a los tokens de entrada interactuar entre sí, y una capa de dot producto que combina estos tokens para producir la salida. En resumen, la atención de los transformadores se puede resumir en la siguiente ecuación:

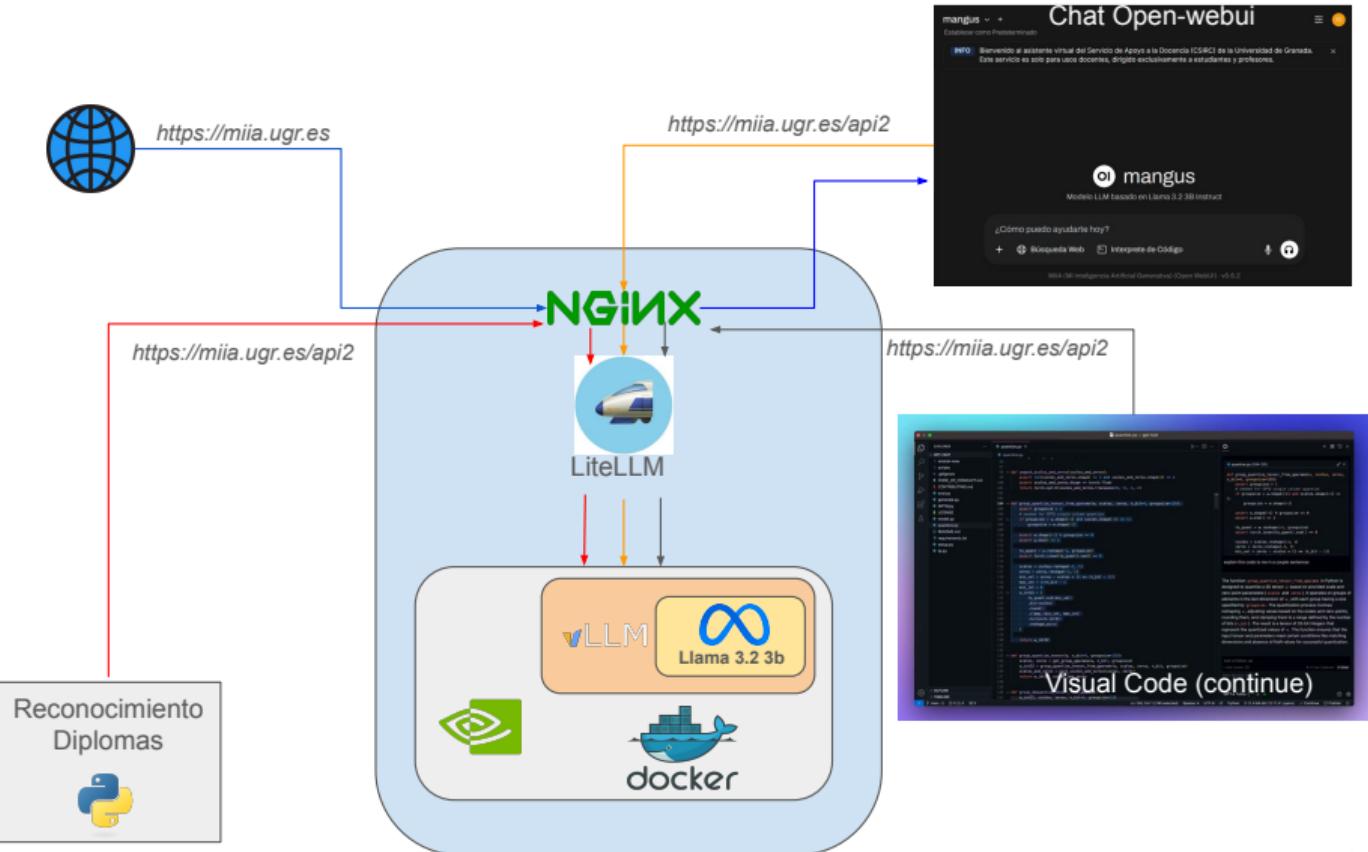
$$\text{Output} = \sum_{i=1}^L \text{Attention}(i) \times \text{Output}_i$$

donde $\text{Attention}(i)$ es la atención dada por el token i y Output_i es el output del token i .



Reconocimiento (Parseo) de Documentos





UNIVERSIDAD
DE GRANADA

Personal

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



Preguntas

Presentación y Código

<https://github.com/isvida/2025-RedIris>



UNIVERSIDAD
DE GRANADA

JT 2025
RedIRIS