

LLM COMO SERVICIO

MiLA 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



CSIRC
CENTRO DE SERVICIOS DE INFORMÁTICA Y REDES DE COMUNICACIONES

Jt 2025
Red IRIS



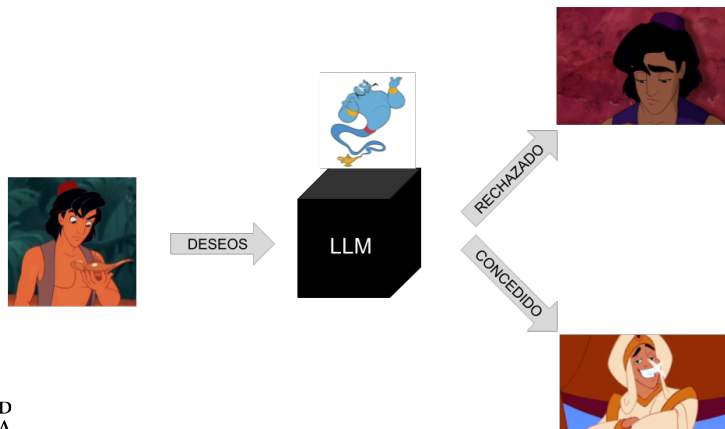
UNIVERSIDAD
DE GRANADA

Jt 2025
Red IRIS

Inteligencia Artificial (según los "Mass Media")

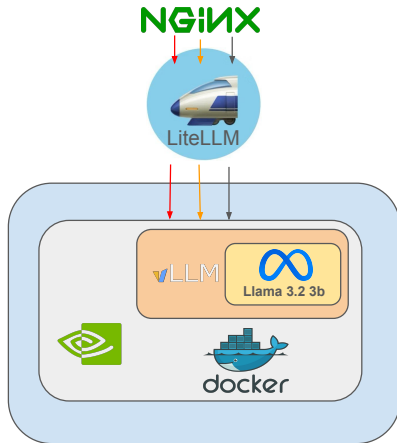


Lámpara Mágica



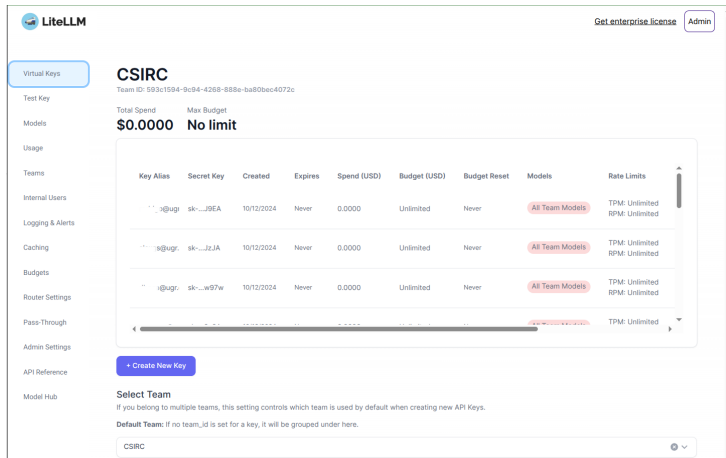
Core MilA

Componentes del Servicio



LiteLLM: proxy LLM

Definición de API keys: Virtual Keys



The screenshot displays the LiteLLM web interface. On the left is a sidebar with navigation links: Virtual Keys (selected), 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' with a team ID '593c1594-9c94-4268-888e-ba80bec4072c'. It shows 'Total Spend: \$0.0000' and 'Max Budget: No limit'. Below this is a table of virtual 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

Below the table is a '+ Create New Key' button. Further down, there is a 'Select Team' section with the text: 'If you belong to multiple teams, this setting controls which team is used by default when creating new API Keys. Default Team: If no team_id is set for a key, it will be grouped under here.' A dropdown menu shows 'CSIRC' as the selected team.

LiteLLM: proxy LLM

Definición de modelos y control de accesos.

LiteLLM [Get enterprise license](#) [Admin](#)

Virtual Keys [All Models](#) [Add Model](#) [/health](#) [Models](#) [Model Analytics](#) [Model Retry Settings](#) Last Refreshed: 20/5/2025, 20:30:29

Test Key

Models

Usage

Teams

Internal Users

Logging & Alerts

Caching

Budgets

Router Settings

Pass-Through

Admin Settings

API Reference

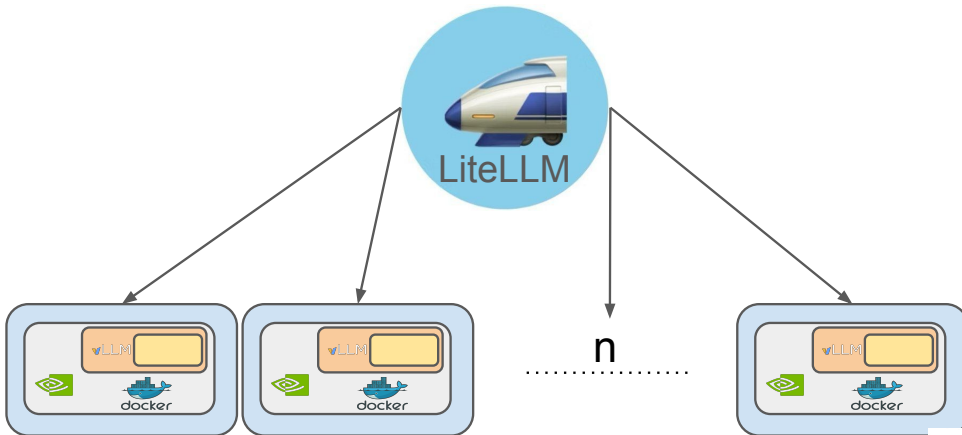
Model Hub

Filter by Public Model Name

Public Model Name	Provider	LiteLLM Model	API Base	Input Price /1M Tokens (\$)	Output Price /1M Tokens (\$)	Created At	Created By	Status
mangus-code2	hosted_vllm	hosted_vllm/Qwen/Qwen...	http://orca1.u...gr.es:	-	-	-	-	DB Model Info Edit Delete
mangus-code	hosted_vllm	hosted_vllm/Qwen/Qwen...	http://orca1.u...gr.es:	-	-	-	-	DB Model Info Edit Delete
mangus	hosted_vllm	hosted_vllm/unsloth/...	http://orca1.u...gr.es:	-	-	-	-	DB Model Info Edit Delete

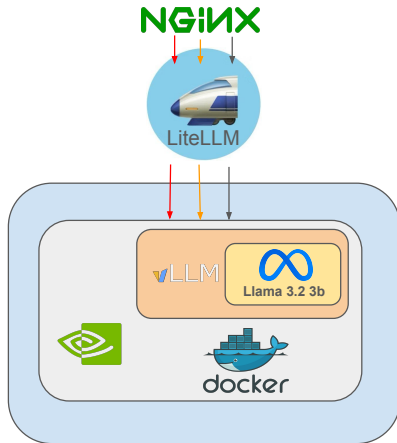
LiteLLM: proxy LLM

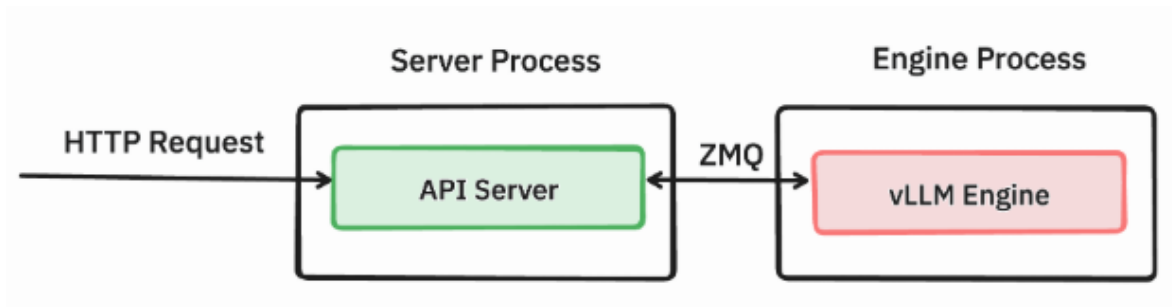
Escalado del servicio LLM



Core MilA

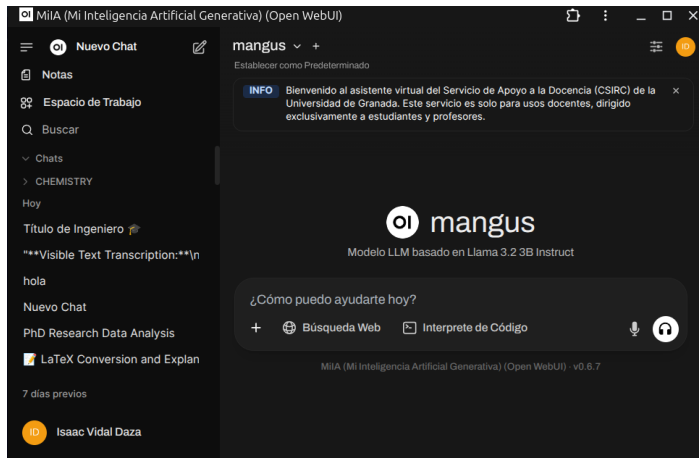
Componentes del Servicio





Servicios Ofrecidos con MiLA Core

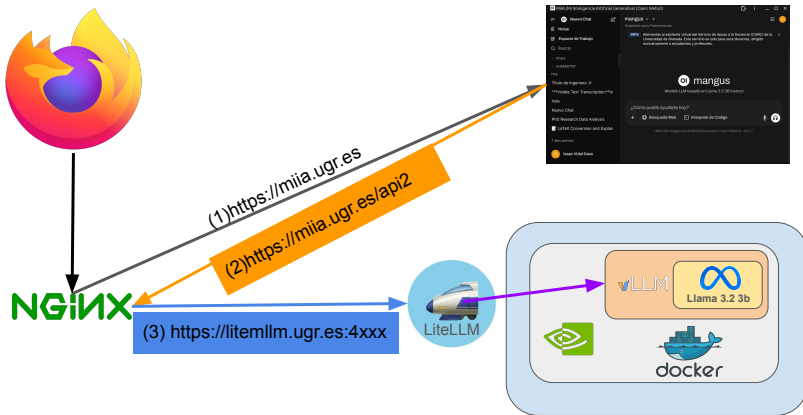
MiLA Chat (*Open WebUI*)



UNIVERSIDAD
DE GRANADA

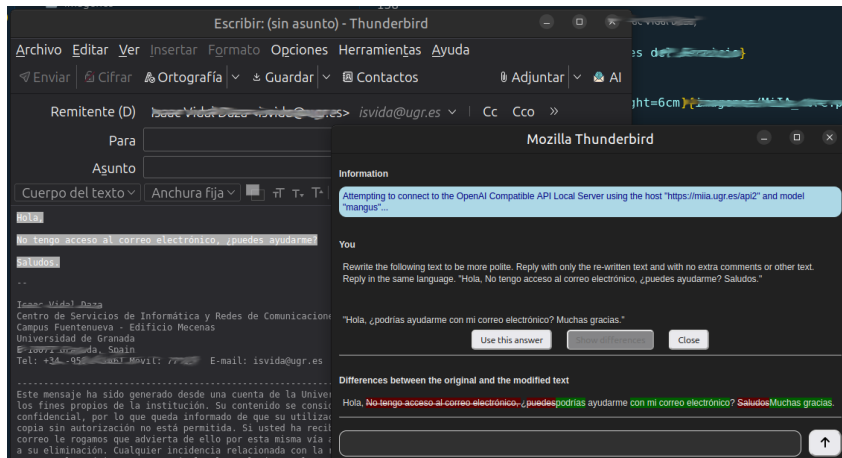
Servicios Ofrecidos con MiLA Core

MiLA Chat (*Open WebUI*): Arquitectura



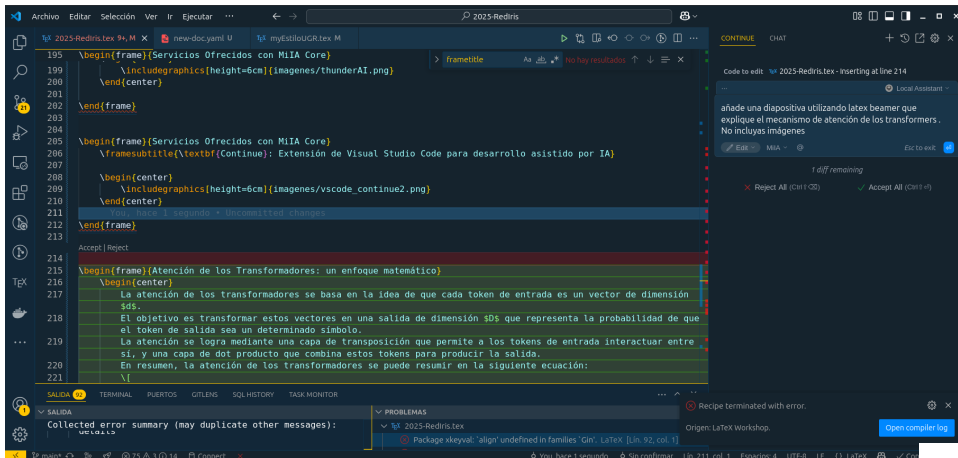
Servicios Ofrecidos con MiLA Core

ThunderAI: Extensión de Thunderbird para correo electrónico



Servicios Ofrecidos con MiIA Core

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



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

Carrera Horizontal del PTGAS

- Entrega de documentación para la evaluación.
- Análisis de 11510 documentos en poco tiempo.



UNIVERSIDAD
DE GRANADA



CENTRO
MEDITERRÁNEO

DÑA. *[REDACTED]* MARTOS *[REDACTED]*, SUBDIRECTORA DEL CENTRO
MEDITERRÁNEO DE LA UNIVERSIDAD DE GRANADA

HACE CONSTAR: Que D. Isaac Vidal Daza, con D.N.I. nº *[REDACTED]*, ha superado, con la calificación de SOBRESALIENTE (9), el Curso "**Ciencia de Datos: Un Enfoque Práctico en la Era del Big Data (4ª edición)**" que, con una duración de 30 horas presenciales y 45 horas no presenciales, se ha celebrado en Granada, del 1 al 23 de febrero de 2018, con las ponencias que a continuación se indican:

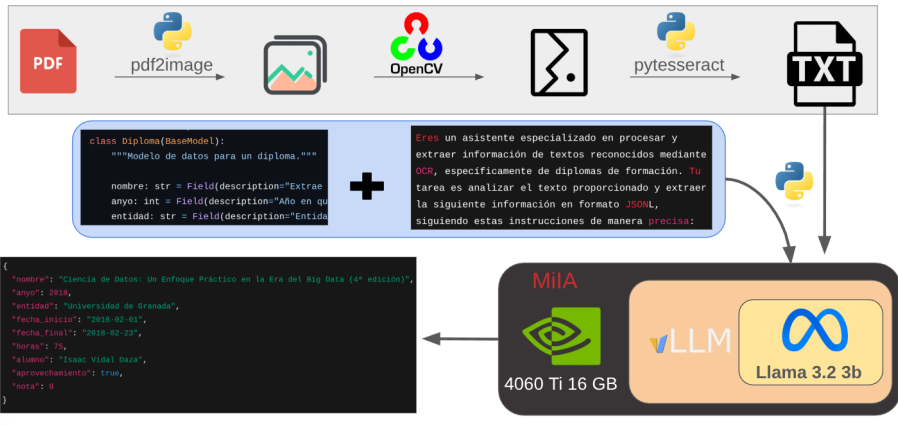
- Tema 1: Ciencia de Datos, analítica avanzada y big data.
- Tema 2: Análisis exploratorio de datos: visualización.
Jorge Casillas Barranquero
- Tema 3: Fundamentos de clasificación: árboles de decisión, lazy, RNA, bayesianos, evaluación.
Salvador García
- Práctica 1: KNIME: predicción fundamental.



UNIVERSIDAD
DE GRANADA

Reconocimiento (*Parseo*) de Documentos

EPFReader





<https://miia.ugr.es>

<https://miia.ugr.es/api2>

NGINX



LiteLLM

LLM

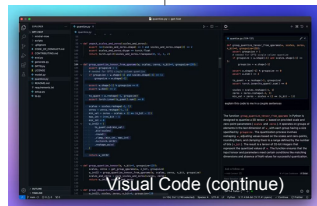
Llama 3.2 3b



Reconocimiento
Diplomas



<https://miia.ugr.es/api2>



UNIVERSIDAD
DE GRANADA

Jt2025
Red
IRIS

Personal

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

Presentación y Código

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