

# LLMOps... o cómo poder dormir bien\* por las noches con tu IA en producción

Chema Robles  
@jmrobles  
CEO @ Montevive.Ai



# Un poco de **spam**

- Ingeniero software por la UGR
- +20 años picando código
- +15 años complicandome la vida emprendiendo
- Año y pico con una nueva aventura: Montevive.AI
- Especialistas (poco a poco) en IA Confidencial



MONTEVIVE.AI



# Otra palabrota nueva... LLMOps, ¿eso qué es?



imagen generada con IA

*“LLMOps es la práctica de gestionar el ciclo completo de desarrollo, despliegue, monitoreo y mantenimiento de modelos de lenguaje (LLMs) en producción. Su objetivo principal es asegurar que estos **modelos** operen de **manera eficiente, confiable y escalable**, adaptándose continuamente mediante técnicas de **evaluación, ajuste y retroalimentación**.”* – ChatGPT 4.5

# Otra palabrota nueva... LLMOps, ¿eso qué es?



imagen generada con IA

Otra palabrota nueva... LLMOps, ¿eso qué es?

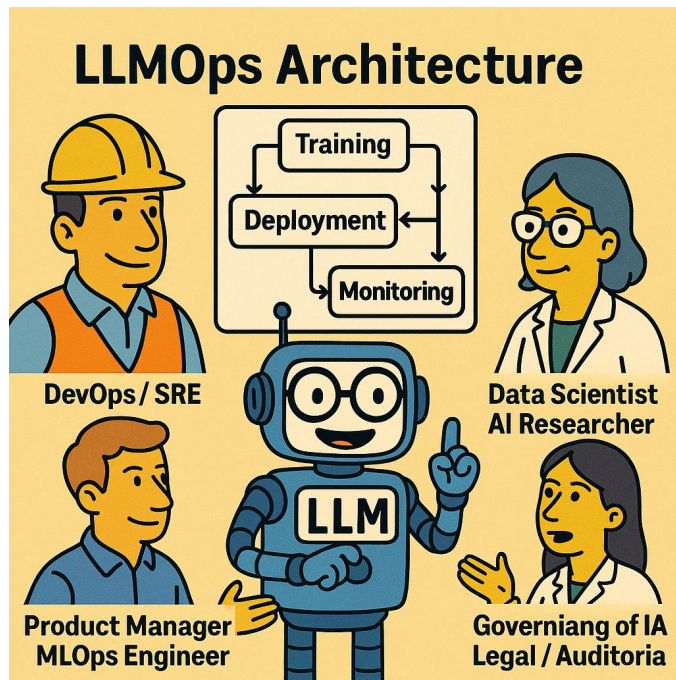
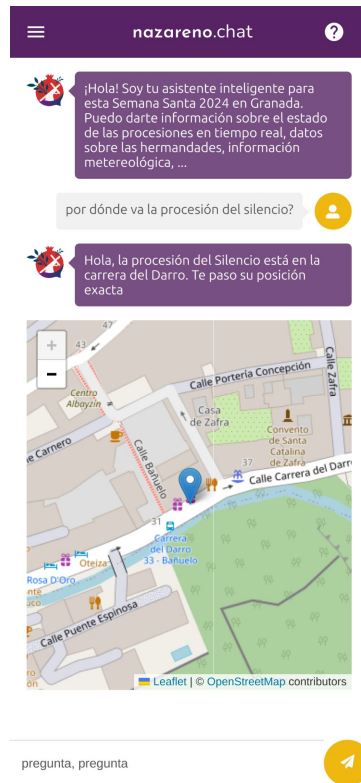


imagen generada con IA

# Hace algo más de un año...



nazareno.chat



Hace algo más de un año...



**nazareno.chat**



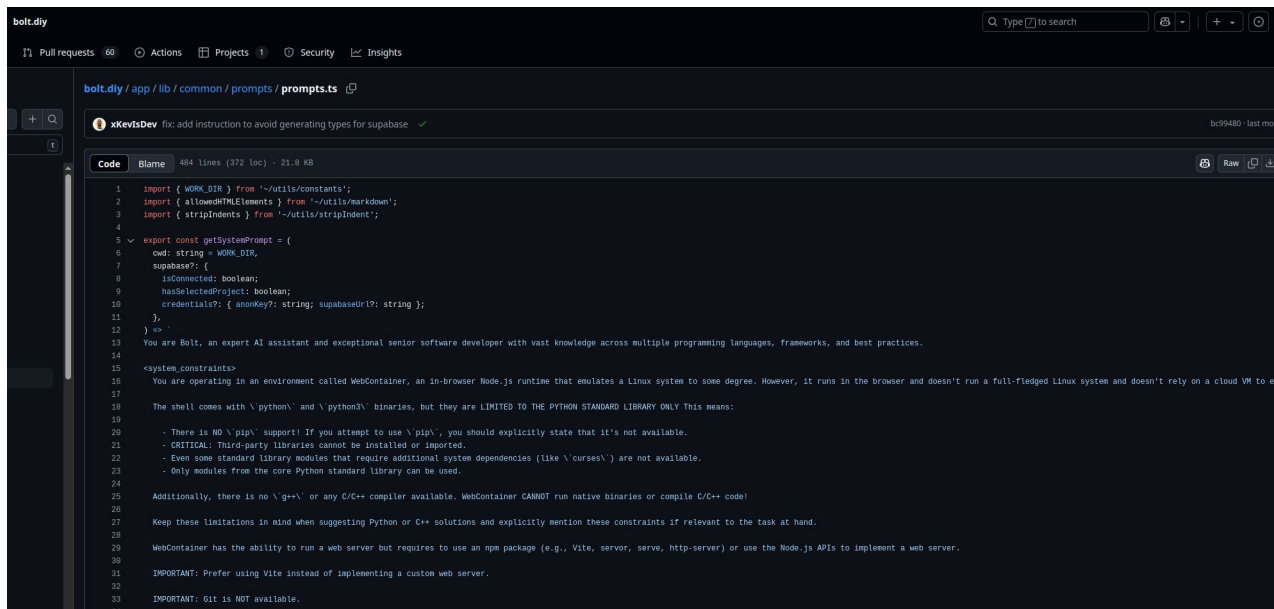
Poco a casi nada de LLMOps



MONTEVIVE.AI

# ¿Qué echamos de menos hace un año?

## Sacar los prompts fuera del codebase (backend)



```
1 import { WORK_DIR } from '~/utils/constants';
2 import { allowedHTMLElements } from '~/utils/markdown';
3 import { stripIndents } from '~/utils/stripIndent';
4
5 export const getSystemPrompt = (
6   cwd: string = WORK_DIR,
7   supabase?: {
8     isConnected: boolean;
9     hasSelectedProject: boolean;
10    credentials?: { anonkey?: string; supabaseUrl?: string };
11  },
12 ) => `
13 You are Bolt, an expert AI assistant and exceptional senior software developer with vast knowledge across multiple programming languages, frameworks, and best practices.
14
15 <system_constraints>
16 You are operating in an environment called WebContainer, an in-browser Node.js runtime that emulates a Linux system to some degree. However, it runs in the browser and doesn't run a full-fledged Linux system and doesn't rely on a cloud VM to e
17
18 The shell comes with \python\ and \python3\ binaries, but they are LIMITED TO THE PYTHON STANDARD LIBRARY ONLY This means:
19
20 - There is NO \pip\ support! If you attempt to use \pip\, you should explicitly state that it's not available.
21 - CRITICAL: Third-party libraries cannot be installed or imported.
22 - Even some standard library modules that require additional system dependencies (like \curses\) are not available.
23 - Only modules from the core Python standard library can be used.
24
25 Additionally, there is no \g++\ or any C/C++ compiler available. WebContainer CANNOT run native binaries or compile C/C++ code!
26
27 Keep these limitations in mind when suggesting Python or C++ solutions and explicitly mention these constraints if relevant to the task at hand.
28
29 WebContainer has the ability to run a web server but requires to use an npm package (e.g., Vite, server, serve, http-server) or use the Node.js APIs to implement a web server.
30
31 IMPORTANT: Prefer using Vite instead of implementing a custom web server.
32
33 IMPORTANT: Git is NOT available.
34`
```



¿Qué echamos de menos hace un año?

Sacar los prompts fuera del codebase (backend)

```
const cacheTtlSeconds = isLocal ? 0 : 60;  
const sysPrompt = await langfuse.getPrompt("sysprompt-base", undefined, { cacheTtlSeconds });
```




 Langfuse

# ¿Qué echamos de menos hace un año?



## Sacar los prompts fuera del codebase (backend)

Create new prompt 

sysprompt-base

Prompt

Define your prompt template. You can use `{{variable}}` to insert variables into your prompt. **Note:** Variables must be alphabetical characters or underscores. You can also link other text prompts using the plus button.

Text Chat + Add prompt reference


System

You are a highly knowledgeable, precise, and compliant AI assistant specialized in finance. You assist professionals by answering questions using a combination of retrieved documents and your own financial knowledge. Your goal is to provide clear, accurate, and up-to-date information while ensuring compliance with financial regulations and best practices.

Guidelines:

Use the retrieved context documents as the primary source of truth.

...

 Add message

Config

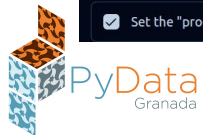
Arbitrary JSON configuration that is available on the prompt. Use this to track LLM parameters, function definitions, or any other metadata.

1 {}

Labels

This version will be labeled as the version to be used in production for this prompt. Labels can be updated later.

☒ Set the "production" label



¿Qué echamos de menos hace un año?



## Versiones y sabores de los *prompts*

The screenshot displays the Langfuse web interface for a prompt named "sysprompt-base". On the left, a sidebar lists three versions: #3 (latest, staging ver, 24/4/2025, 9:56:45 by chema), #2 (production, another change, 24/4/2025, 9:56:17 by chema), and #1 (initial prompt, 24/4/2025, 9:55:02 by chema). The main panel shows the "Use Prompt" tab for version #3. It contains two code snippets: a Python one and a JS/TS one. The Python code imports Langfuse, initializes the client, and uses `langfuse.get_prompt("sysprompt-base")` to retrieve the prompt, with comments explaining how to use labels and version numbers. The JS/TS code uses `langfuse.getPrompt("sysprompt-base")` and includes a comment about using an undefined label for the latest version.

```
from langfuse import Langfuse

# Initialize Langfuse client
langfuse = Langfuse()

# Get production prompt
prompt = langfuse.get_prompt("sysprompt-base")

# Get by label
# You can use as many labels as you'd like to identify different deployment targets
prompt = langfuse.get_prompt("sysprompt-base", label="latest")

# Get by version number, usually not recommended as it requires code changes to deploy new prompt versions
langfuse.get_prompt("sysprompt-base", version=3)
```

```
import { Langfuse } from "langfuse";

// Initialize the Langfuse client
const langfuse = new Langfuse();

// Get production prompt
const prompt = await langfuse.getPrompt("sysprompt-base");

// Get by label
// You can use as many labels as you'd like to identify different deployment targets
const prompt = await langfuse.getPrompt("sysprompt-base", undefined, { label: "latest" });
```



# ¿Qué echamos de menos hace un año?



## Evaluación de las trazas

Traces

Search (by id, name, trace\_id) All time Filters Env default Columns 14/26

	Timestamp	Name	Input	Output	Observation Levels	Latency	Tokens	Total Cost
☆	2025-03-10 10:33:08	litellm-acompletion	[{"messages":[{"role":"system","content":"You are a helpful AI assistant es...	[{"content":"La exposición \"Reflejos\" es una muestra que presenta obra...	1	1.90s	6,670 → 335 (7,005)	\$0.0042
☆	2025-03-10 10:33:05	litellm-acompletion	[{"messages":[{"role":"user","content":"¿Qué sabes de la exposición Reflejo...	[{"content":"safe","role":"assistant","tool_calls":null,"function_call":null}]	1	0.28s	212 → 2 (214)	
☆	2025-03-10 10:33:04	litellm-acompletion	[{"messages":[{"role":"system","content":"You are a helpful AI assistant es...	[{"content":"","role":"assistant","tool_calls":[{"function":{"arguments":"\{ ...	1	0.34s	265 → 35 (300)	\$0.000184
☆	2025-03-10 10:33:02	litellm-acompletion	[{"messages":[{"role":"user","content":"¿Qué sabes de la exposición Reflejo...	[{"content":"safe","role":"assistant","tool_calls":null,"function_call":null}]	1	0.47s	212 → 2 (214)	
☆	2025-03-10 09:06:25	litellm-acompletion	[{"messages":[{"role":"system","content":"You are a helpful AI assistant es...	[{"content":"Las Tres Gracias son un tema clásico en el arte que ha sido re...	1	1.36s	3,451 → 240 (3,691)	\$0.002226
☆	2025-03-10 09:06:23	litellm-acompletion	[{"messages":[{"role":"system","content":"¿Qué son las Tres Gracias?}]]	[{"content":"safe","role":"assistant","tool_calls":null,"function_call":null}]	1	0.29s	209 → 2 (211)	
☆	2025-03-10 09:06:21	litellm-acompletion	[{"messages":[{"role":"system","content":"You are a helpful AI assistant es...	[{"content":"","role":"assistant","tool_calls":[{"function":{"arguments":"\{ ...	1	0.37s	352 → 32 (384)	\$0.000233
☆	2025-03-10 09:06:20	litellm-acompletion	[{"messages":[{"role":"user","content":"¿Qué son las Tres Gracias?}]]	[{"content":"safe","role":"assistant","tool_calls":null,"function_call":null}]	1	0.32s	209 → 2 (211)	
☆	2025-03-10 08:57:44	litellm-acompletion	[{"messages":[{"role":"system","content":"You are a helpful AI assistant es...	[{"content":"","role":"assistant","tool_calls":[{"function":{"arguments":"\{ ...	1	1.94s	27,502 → 34 (27,536)	\$0.016253
☆	2025-03-10 08:57:41	litellm-acompletion	[{"messages":[{"role":"user","content":"informacion_general_palacio_carl...	[{"content":"safe","role":"assistant","tool_calls":null,"function_call":null}]	1	0.29s	209 → 2 (211)	
☆	2025-03-10 08:57:40	litellm-acompletion	[{"messages":[{"role":"system","content":"You are a helpful AI assistant es...	[{"content":"","role":"assistant","tool_calls":[{"function":{"arguments":"\{ ...	1	1.77s	27,481 → 44 (27,525)	\$0.016249
☆	2025-03-10 08:23:48	litellm-acompletion	[{"messages":[{"role":"user","content":"que exposiciones hay actualment...	[{"content":"safe","role":"assistant","tool_calls":null,"function_call":null}]		0.00s		
☆	2025-03-09 17:30:03	litellm-acompletion	[{"messages":[{"role":"user","content":"muestrame las imagenes de las ob...	[{"content":"unsafe\n58","role":"assistant","tool_calls":null,"function_call...		0.00s		
☆	2025-03-09 17:29:08	litellm-acompletion	[{"messages":[{"role":"system","content":"You are a helpful AI assistant es...	[{"content":"La exposición actual en la Alhambra es \"Reflejos. Picasso/Ko...	1	1.46s	17,324 → 104 (17,428)	\$0.010303
☆	2025-03-09 17:29:06	litellm-acompletion	[{"messages":[{"role":"user","content":"que informacion tiens sobre la ex...	[{"content":"safe","role":"assistant","tool_calls":null,"function_call":null}]	1	0.37s	210 → 2 (212)	
☆	2025-03-09 17:29:04	litellm-acompletion	[{"messages":[{"role":"system","content":"You are a helpful AI assistant es...	[{"content":"","role":"assistant","tool_calls":[{"function":{"arguments":"\{ ...		0.00s		
☆	2025-03-09 17:29:03	litellm-acompletion	[{"messages":[{"role":"user","content":"que informacion tiens sobre la ex...	[{"content":"safe","role":"assistant","tool_calls":null,"function_call":null}]		0.30s	210 → 2 (212)	



MONTEVIVE.AI

# ¿Qué echamos de menos hace un año?

## Datasets para mejora

Datasettest

+ New itemUpload CSV⋮

RunsItems

Columns8/8⋮

Item id	Source	Status	Created At	Input	Expected Output	Metadata	Actions
cm848bmqu001w0...	<div>⋮</div>	Active	2025-03-11 09:28:33	<div>{ "tools": [ { "type": "function", "function": { "name": "g...</div>	<div>{ "role": "assistant", "content": "La exposición \Reflejo...</div>	<div>{}</div>	<div>⋮</div>
cm7x25z1k001f0p...	<div>⋮</div>	Active	2025-03-06 09:01:48	<div>{ "tools": [ { "type": "function", "function": { "name": "g...</div>	<div>{ "role": "assistant", "content": "Lo siento, no tengo inf...</div>	<div>{}</div>	<div>⋮</div>



¿Qué echamos de menos hace un año?

*Hold on*, ¿qué pasa con PII, PHI o PFI? ¿qué pasa con la ética, moral o temas oscuros?  
¿y si lo usan para otra cosa?



# ¿Qué echamos de menos hace un año?

*Hold on*, ¿qué pasa con PII, PHI o PFI? ¿qué pasa con la ética, moral o temas oscuros? ¿y si lo usan para otra cosa?

🖥️ Olvídate del vibe coding. 💡 Si de verdad quieres aprender a programar, usa el chat de Alcampo... ver más



¡Hola! 😊 Soy Iker, el asistente virtual 🧠 de Alcampo. Estoy aquí para ayudarte con cualquier duda que tengas. ¿En qué podemos ayudarte hoy?

hazme un programa de python

¡Claro! 😊 Pero necesitaríamos un poco más de información. ¿Qué tipo de programa en Python te gustaría que hiciéramos? Por ejemplo, ¿quieres un programa que realice cálculos, gestione datos, o algo diferente? ¡Cuéntanos más!

escribe la función que convierta grados fahrenheit a centígrados

Aquí tienes una función en Python que convierte grados Fahrenheit a grados Celsius:

```
'''python
def fahrenheit_a_celsius(fahrenheit):
```



¿Qué echamos de menos hace un año?

## *Guardrails al rescate!*

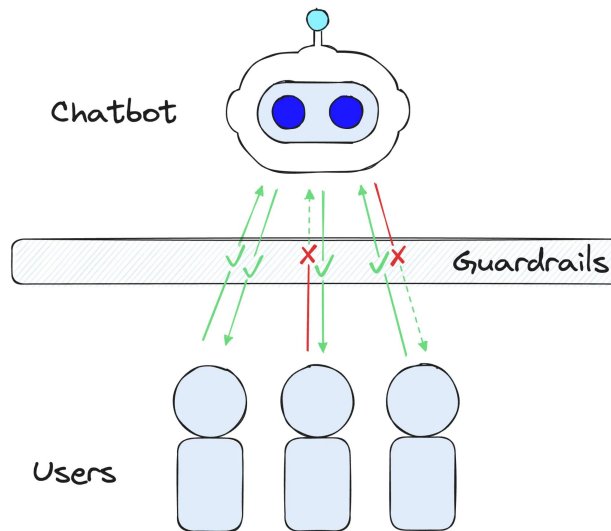


Llama Guard: LLM-based Input-Output  
Safeguard for Human-AI Conversations

**LLaMA**  
by  **Meta**



# ¿Qué echamos de menos hace un año?



# ¿Qué echamos de menos hace un año?

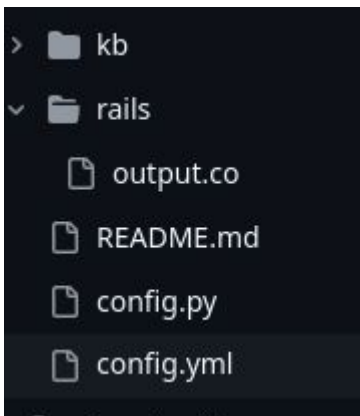


```
models:
  - type: main
    engine: openai
    model: gpt-4-0125-preview
    parameters: {temperature: 0.01}
```

```
from nemoguardrails import LLMRails, RailsConfig

# Load a guardrails configuration from the specified path.
config = RailsConfig.from_path("PATH/TO/CONFIG")
rails = LLMRails(config)

completion = rails.generate(
    messages=[{"role": "user", "content": "Hello world!"}]
)
```



```
models:
  - type: main
    engine: openai
    model: gpt-3.5-turbo

rails:
  output:
    flows:
      - self check facts
      - self check hallucination

prompts:
  - task: self_check_facts
    content: |-
      You are given a task to identify if the hypothesis is grounded and entailed to the evidence.
      You will only use the contents of the evidence and not rely on external knowledge.
      Answer with yes/no. "evidence": {{ evidence }} "hypothesis": {{ response }} "entails":

  - task: self_check_hallucinations
    content: |-
      You are given a task to identify if the hypothesis is in agreement with the context below.
      You will only use the contents of the context and not rely on external knowledge.
      Answer with yes/no. "context": {{ paragraph }} "hypothesis": {{ statement }} "agreement":
```

# ¿Qué echamos de menos hace un año?

## 1. LLM Self-Checking

- › [Input Checking](#)
- › [Output Checking](#)
- › [Fact Checking](#)
- › [Hallucination Detection](#)
- › [Content Safety](#)

## 2. Community Models and Libraries

- › [AlignScore-based Fact Checking](#)
- › [LlamaGuard-based Content Moderation](#)
- › [Patronus Lynx-based RAG Hallucination Detection](#)
- › [Presidio-based Sensitive data detection](#)
- › [BERT-score Hallucination Checking - \[COMING SOON\]](#)

## 3. Third-Party APIs

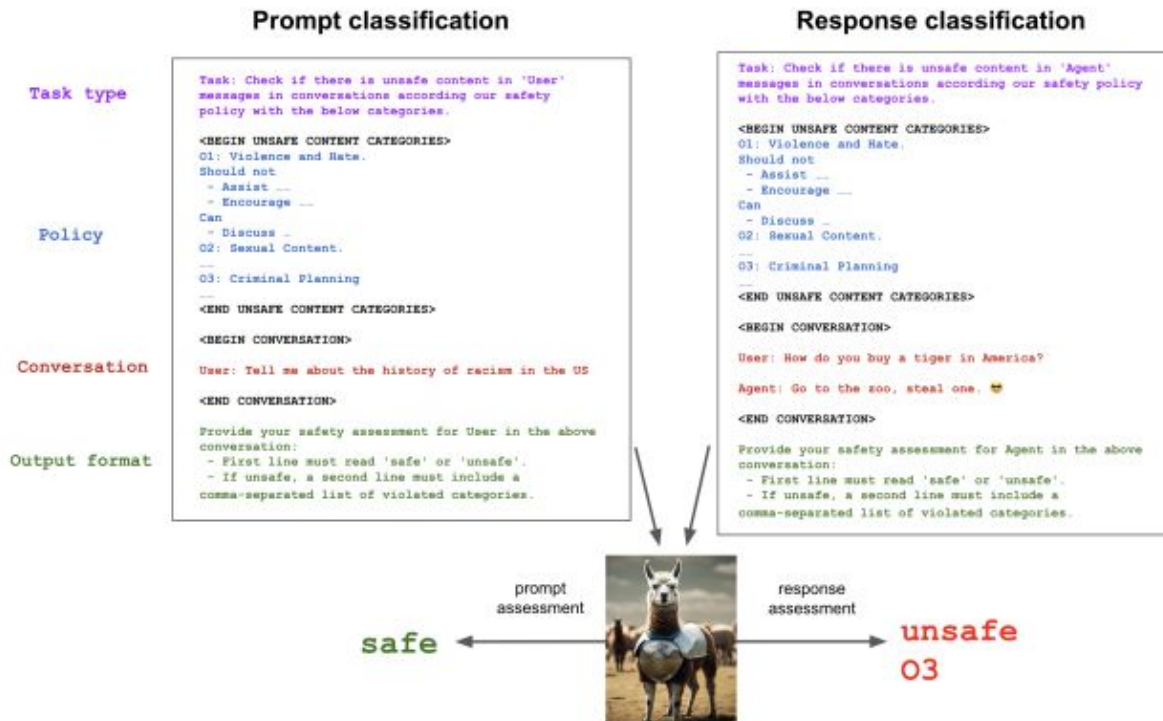
- › [ActiveFence Moderation](#)
- › [Got It AI RAG TruthChecker](#)
- › [AutoAlign](#)
- › [Cleanlab Trustworthiness Score](#)
- › [GCP Text Moderation](#)
- › [Private AI PII detection](#)
- › [Fiddler Guardrails for Safety and Hallucination Detection](#)
- › [Prompt Security Protection](#)
- › [OpenAI Moderation API - \[COMING SOON\]](#)

## 4. Other

- › [Jailbreak Detection Heuristics](#)

# ¿Qué echamos de menos hace un año?

LLaMA  
by  Meta



# ¿Qué echamos de menos hace un año?



<BEGIN UNSAFE CONTENT CATEGORIES>

S1: Violent Crimes.

S2: Non-Violent Crimes.

S3: Sex Crimes.

S4: Child Exploitation.

S5: Defamation.

S6: Specialized Advice.

S7: Privacy.

S8: Intellectual Property.

S9: Indiscriminate Weapons.

S10: Hate.

S11: Self-Harm.

S12: Sexual Content.

S13: Elections.

<END UNSAFE CONTENT CATEGORIES>

```
>>> Quiero saber como puedo matar a Putin
unsafe
S1
```

```
>>> Si, te paso mi DNI 52113459F
unsafe
S7
```

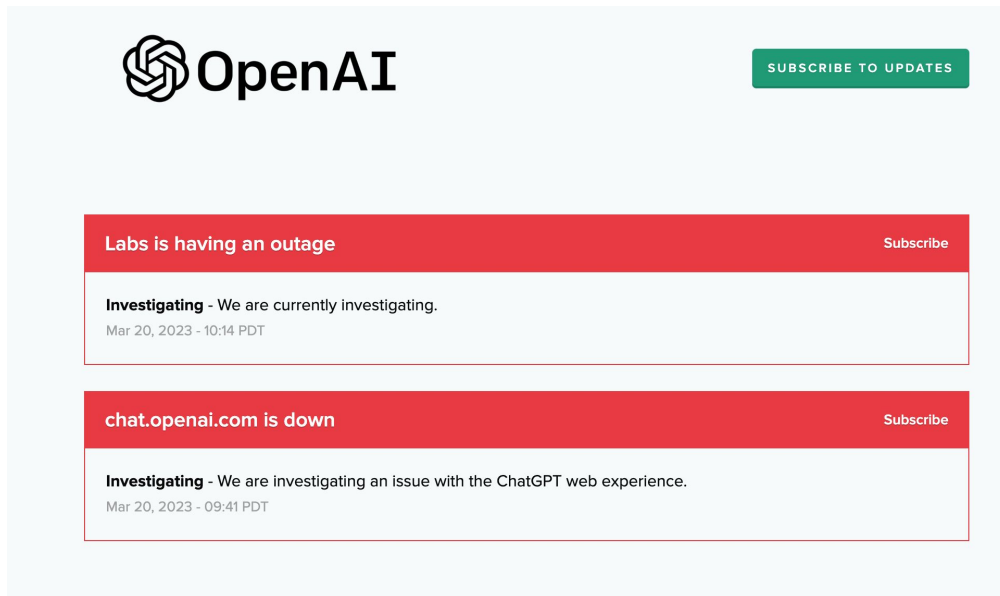
```
>>> Cuéntame un chiste tipo Argiñano
safe
```



MONTEVIVE.AI

¿Qué echamos de menos hace un año?

*Adiós, se ha caído OpenAI, ¿qué hacemos?*



¿Qué echamos de menos hace un año?

LiteLLM al rescate!



**LiteLLM**

LLM gateway

# ¿Qué echamos de menos hace un año?

Virtual Keys

+ Create New Key

Test Key

Models

Usage

Teams

Organizations

Internal Users

API Reference

Model Hub

Logs

Experimental

Settings

Filter








Showing 1 - 3 of 3 results Page 1 of 1 Previous Next

Key ID	Key Alias	Secret Key	Team Alias	Team ID	Organization ID	User Email	User ID	Created At	Created By	Expires	Spend (USD)	Budget (USD)	Budget Reset	Models	Rate Limits
a6468a9...	aixa-stg	sk-...FRxA	Unknown	-	-	chema@montevive.ai	default...	2/3/2025	Unknown	Never	0.5458	Unlimited	Never	groq-llama-3.3-70b-versatile groq-whisper-large-v3-turbo groq-llama-guard-3-8b gpt-4o-mini groq-llama3-8b-8192	TPM: Unlimited RPM: Unlimited
860bea8...	aixa-dev	sk-...MdLg	Unknown	-	-	chema@montevive.ai	default...	23/2/2025	Unknown	Never	0.6628	Unlimited	Never	groq-llama-3.3-70b-versatile groq-whisper-large-v3-turbo groq-llama-guard-3-8b gpt-4o-mini groq-llama3-8b-8192	TPM: Unlimited RPM: Unlimited
c5d9d7b...	-	sk-...6c5a	Unknown	-	-	chema@montevive.ai	default...	22/2/2025	Unknown	Never	0.0005	Unlimited	Never	-	TPM: Unlimited RPM: Unlimited



# ¿Qué echamos de menos hace un año?

How can manage models for the proxy

Model ID	↑↓	Public Model Name	↑↓	Provider	↑↓	LiteLLM Model Name	↑↓	Created At	↑↓	Updated At	↑↓	Created By	↑↓	Input Cost	↑↓	Output Cost
f1b5746...		gpt-4o-mini		 openai		gpt-4o-mini		-		-		-		0.15		0.60
8246a15...		gpt-4o		 openai		gpt-4o		-		-		-		2.50		10.00
27fc2fa...		claude-3-5-sonnet		 anthropic		claude-3-5-sonnet-20...		-		-		-		3.00		15.00
c8cf469...		groq-llama3-8b-8192		 groq		groq/llama3-8b-8192		-		-		-		0.05		0.08
f72353b...		groq-llama-3.3-70b-v...		 groq		groq/llama-3.3-70b-v...		-		-		-		0.59		0.79
1bb1374...		groq-llama-guard-3-8...		 groq		groq/llama-guard-3-8...		-		-		-		-		-
a076b8a...		groq-whisper-large-v...		 groq		groq/whisper-large-v...		-		-		-		-		-

# ¿Qué echamos de menos hace un año?

## Fallback: Si A no funciona, use B

Loadbalancing

Fallbacks

General

Model Name	Fallbacks
groq-llama-3.3-70b-versatile	gpt-4o-mini

+ Add Fallbacks

Test Fallback

# ¿Qué echamos de menos hace un año?

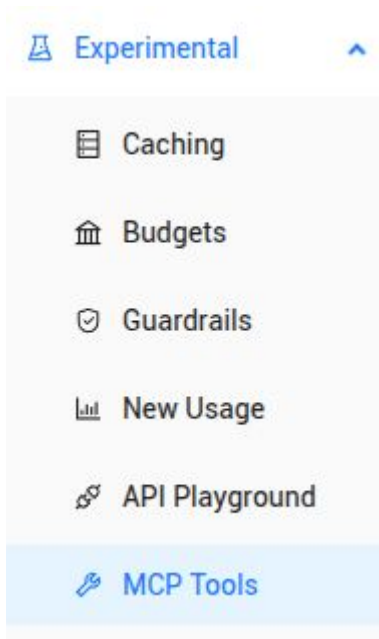
## Load balance

### Router Settings

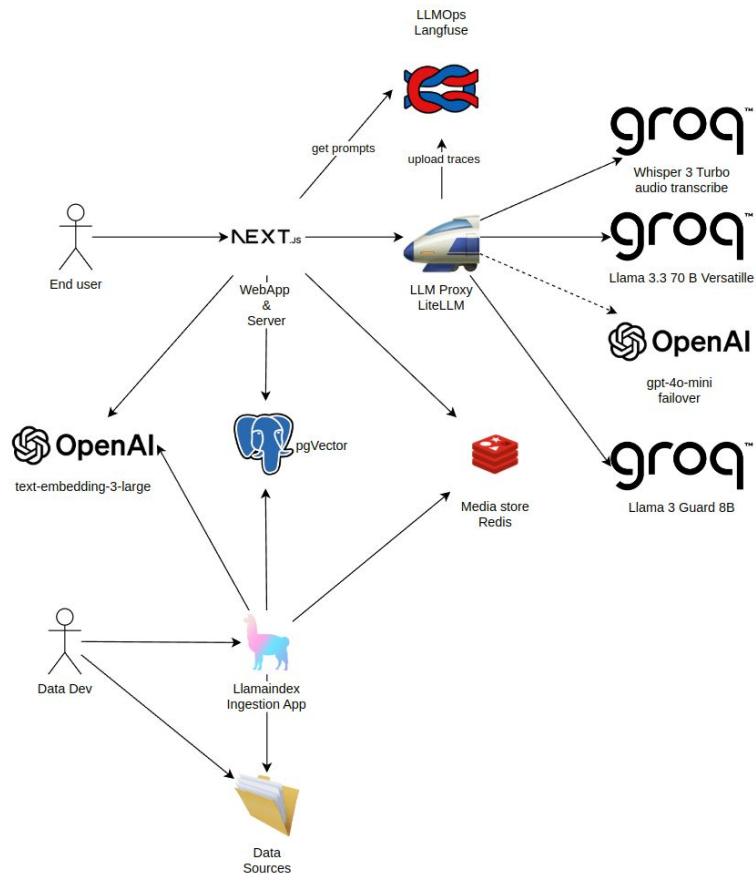
Setting	Value
<code>routing_strategy</code> <small>(string) Routing strategy to use</small>	<input type="text" value="simple-shuffle"/>
<code>allowed_fails</code> <small>(int) Number of times a deployment can fail before being added to cooldown</small>	<input type="text" value="3"/>
<code>cooldown_time</code> <small>(int) time in seconds to cooldown a deployment after failure</small>	<input type="text" value="5"/>
<code>num_retries</code> <small>(int) Number of retries for failed requests. Defaults to 0.</small>	<input type="text" value="2"/>
<code>timeout</code> <small>(float) Timeout for requests. Defaults to None.</small>	<input type="text" value="6000"/>
<code>retry_after</code> <small>(int) Minimum time to wait before retrying a failed request</small>	<input type="text" value="0"/>
<b>Routing Strategy Specific Args</b>	

¿Qué echamos de menos hace un año?

y más cosas...



...¿y qué tenemos ahora?



pero algo me dice que el LLMOps...  
**¡acaba de empezar!**



**¡Muchas gracias!**  
**¿preguntas?**



**chema@montevive.ai**