

# Sistemas Multiagentes

**Andrés Díaz Pace**

**1 & 2 / Nov / 2024**

[andres.diazpace@isistan.unicen.edu.ar](mailto:andres.diazpace@isistan.unicen.edu.ar)

[andres.diazpace@globant.com](mailto:andres.diazpace@globant.com)



**{GEERS}**

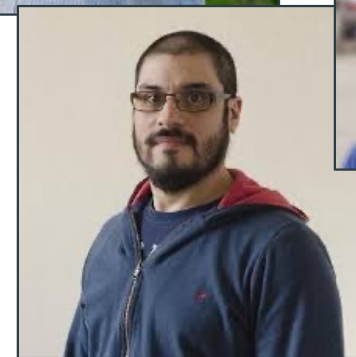
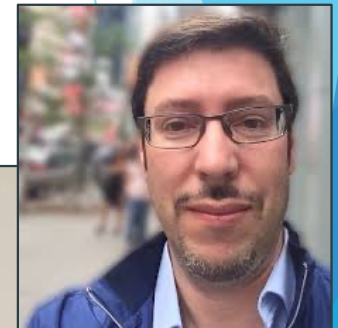
# Agenda

- LLMs y Desarrollo de Aplicaciones
- Patrones basados en LLMs
- Agents & Agentic workflows
  - Memoria, tools, razonamiento
  - Agente ReAct
  - Orquestación de agentes
- Niveles de autonomía
- Lecciones aprendidas y desafíos a futuro

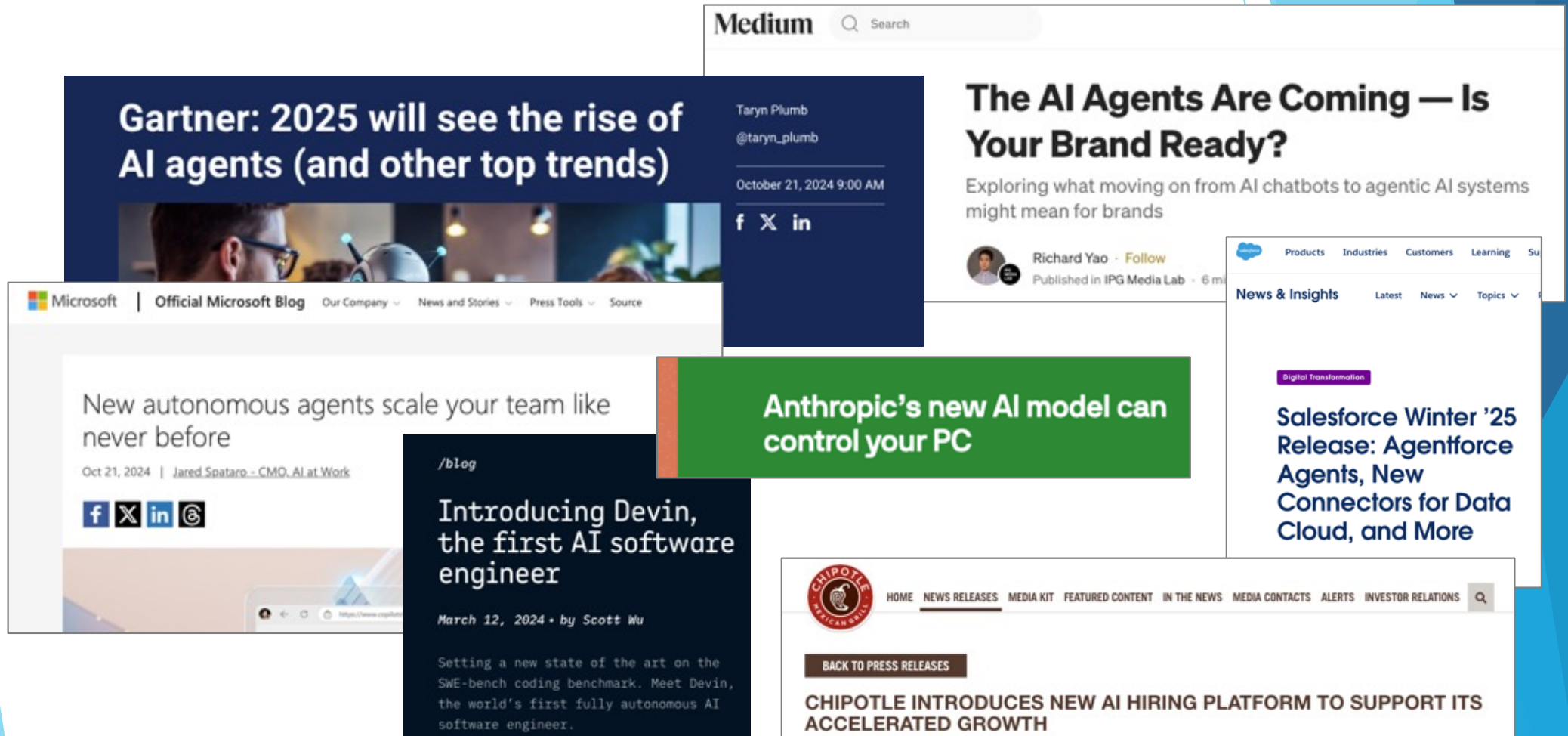


# GEERS.AI 2023 → 2024

- Seguir motorizando la iniciativa (ex-SofIA)
- Introducción a LLMs y posibles aplicaciones
- Diferencias con AI/ML “tradicional”
- **Recuperación con búsqueda semántica (embeddings)**
- **Retrieval Augmented Generation (RAG)**

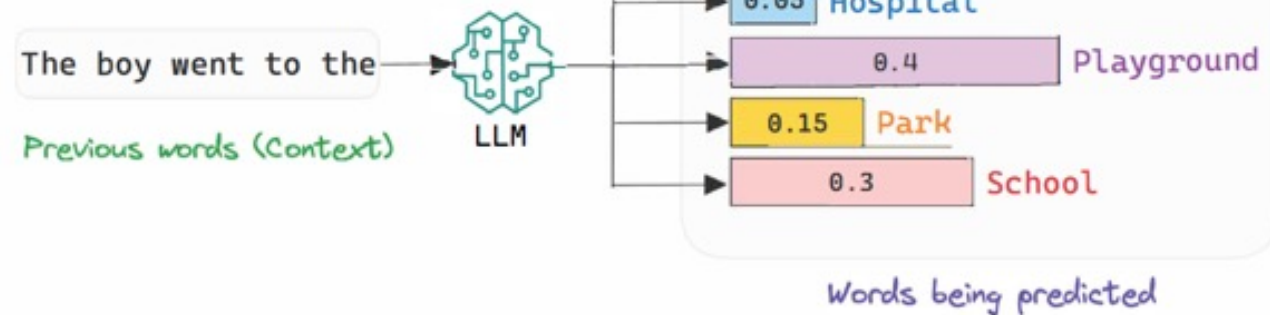


# Agentes en las Noticias



# Contexto: LLMs en breve

- Un Large Language Model puede verse como un **modelo probabilístico sofisticado**
  - Dada una entrada, un LLM genera salidas, en base a patrones que aprende de sus datos de entrenamiento
- El **prompting** es un mecanismo para **condicionar** las salidas del modelo probabilístico, ya sea mediante instrucciones/contexto que hacen que se generen salidas alternativas



# Contexto: Prompting

# Prompt 1

Tell me about: Apple

# Prompt 2

Tell me about: Apple fruit

# Prompt 3

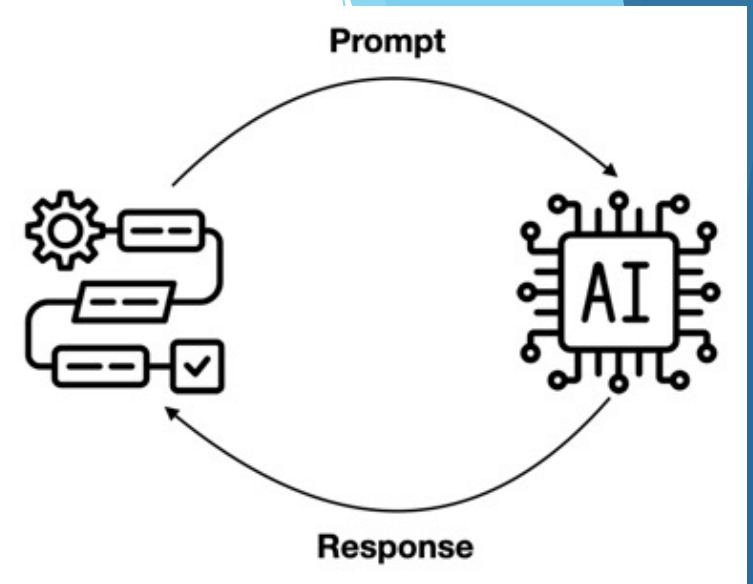
Tell me about: Apple of my eye

# Prompt 4

You are a preschool teacher. Explain how attention in LLMs works.

# Prompt 5

You are an NLP professor. Explain how attention in LLMs works.

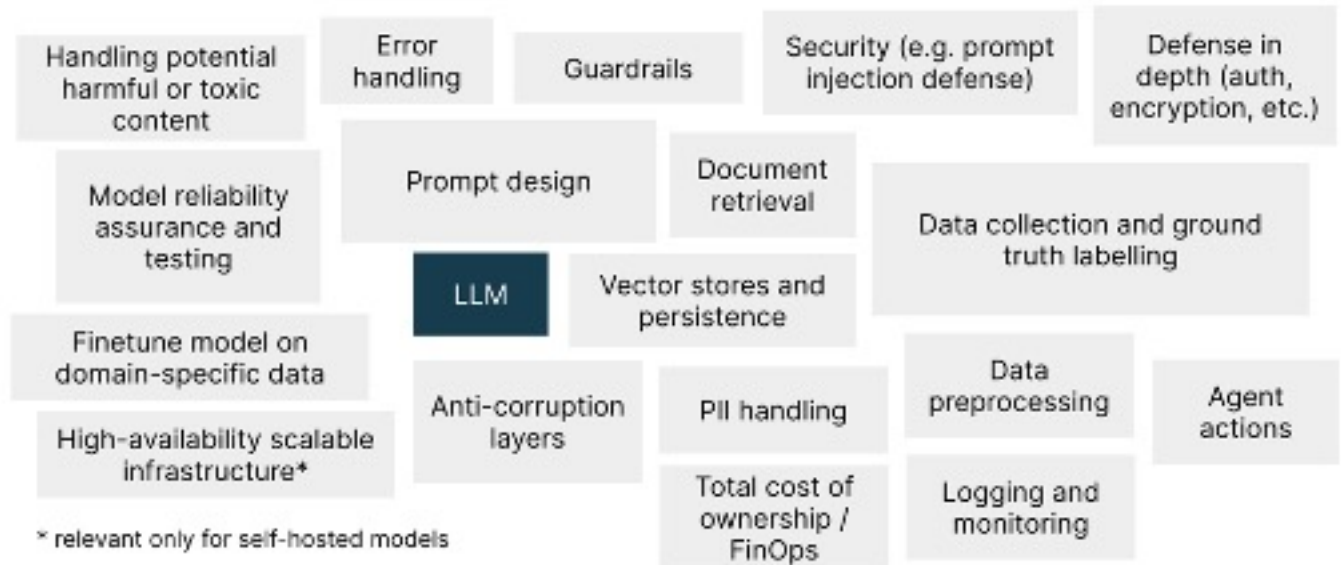
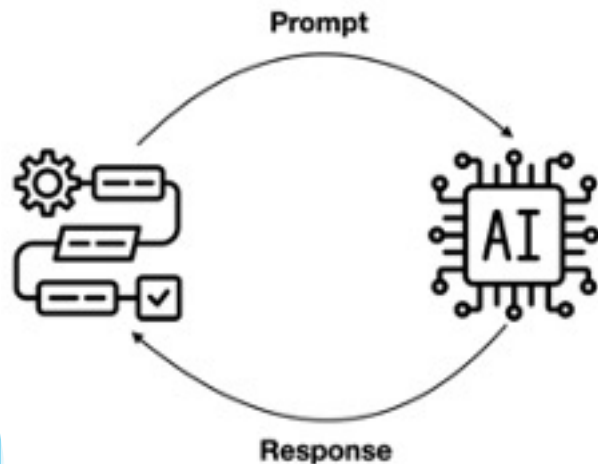


# Contexto: Ingeniería de LLMs

- El objetivo es construir aplicaciones de software soportadas/potenciadas por LLMs (por ej., chatbots, copilots, etc.)

## Architecting LLM applications

The language model is just one part of the technical architecture



\* relevant only for self-hosted models

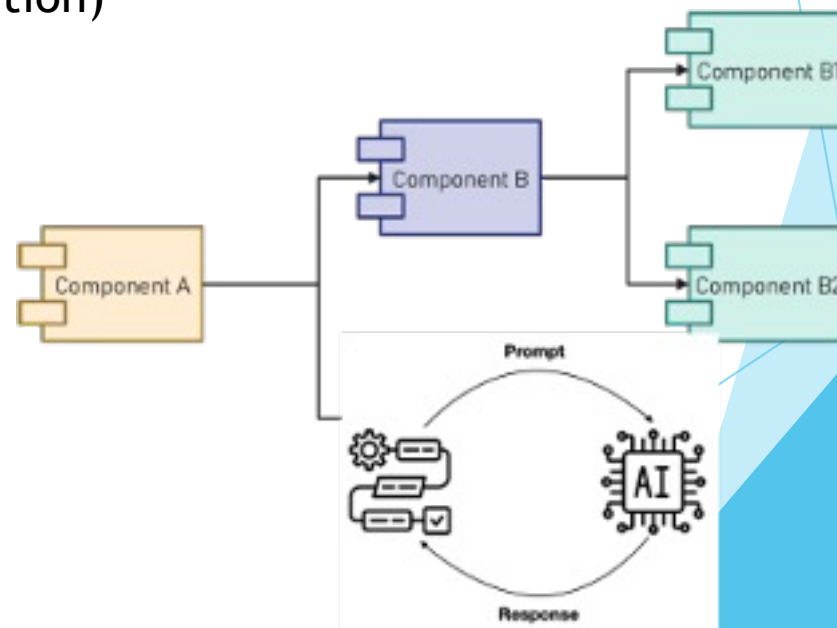
Adapted from: [Machine Learning: The High Interest Credit Card of Technical Debt \(Google\)](https://martinfowler.com/articles/engineering-practices-llm.html)

<https://martinfowler.com/articles/engineering-practices-llm.html>



# Algunos patrones basados en LLMs

1. Prompting
2. Fine-tuning
3. RAG (Retrieval Augmented Generation)
4. Agentic workflow y Agentes





# Volviendo a la base: Prompting

- Conjunto de instrucciones que le dicen a un LLM cómo proceder
- **Variantes:** zero-shot, few-shot, etc.
- Más o menos sofisticado respecto a la estrategia de razonamiento
  - CoT, etc.
- Útil, pero limitado como patrón (desde un punto de vista ingenieril)
- Nota: Algunas técnicas de prompting requieren mecanismos adicionales  
→ memoria

## Chain-of-Thought Prompting

### Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

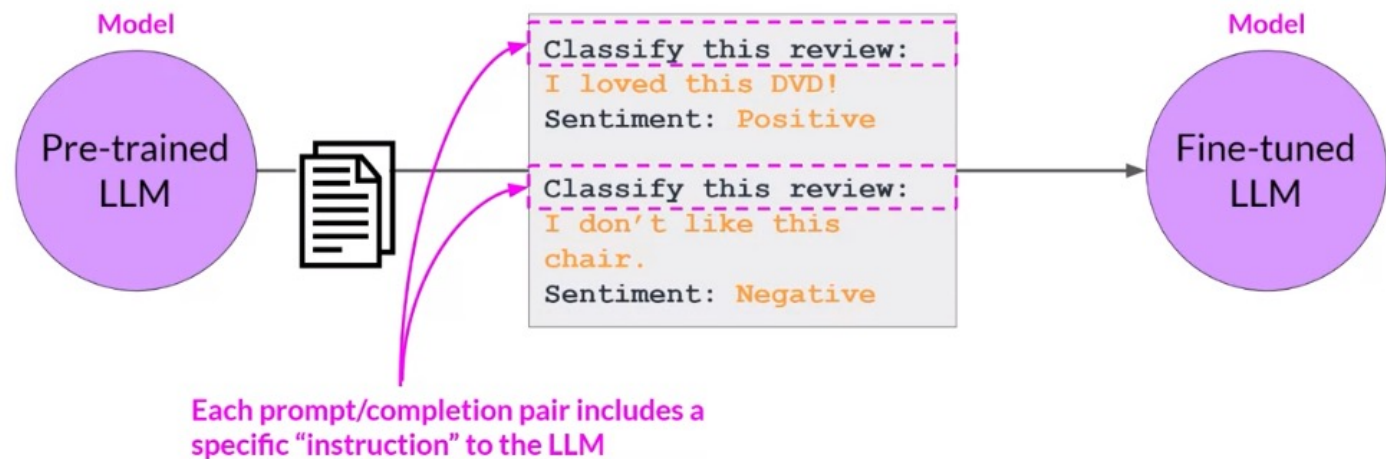
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✓

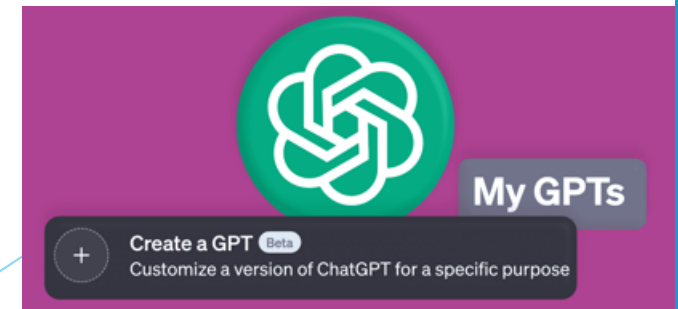
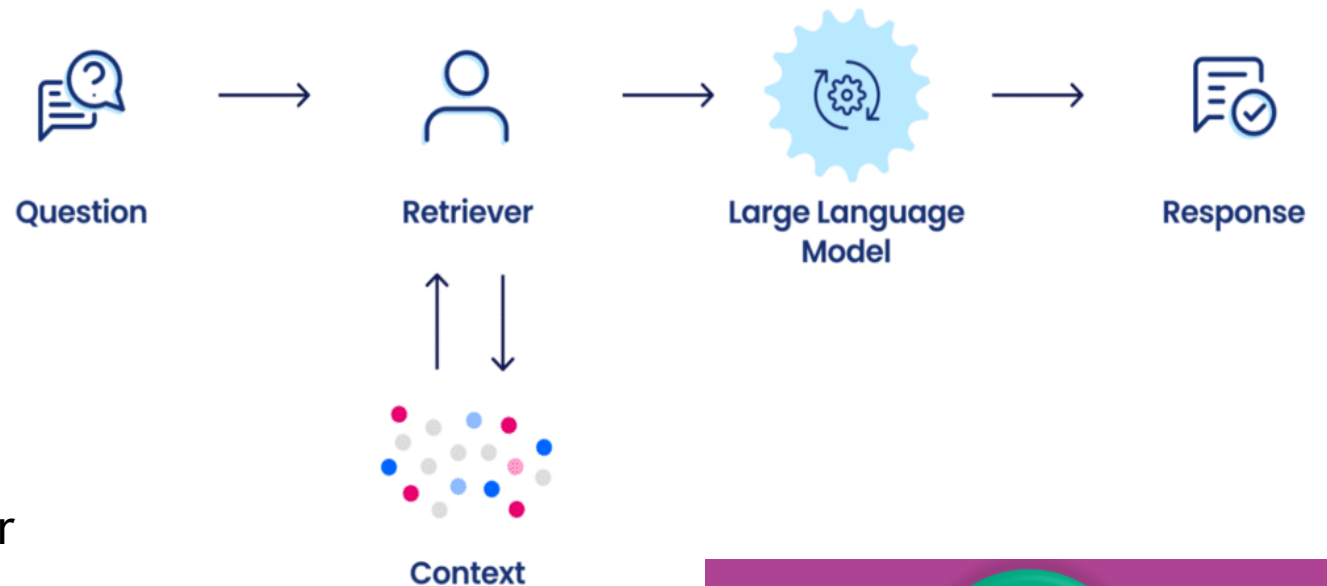
# Especialización: Fine-tuning

- Se busca entrenar/especializar a un LLM para realizar tareas específicas (por ej., en un dominio determinado)
- Requiere contar con datos! (pares pregunta-respuesta)
- Pueden funcionar bien, si se cuenta con datos adecuados, si bien hay que tener cuidado de no sobre-especializar el LLM
- Técnicas más sofisticadas, que permiten “direccionar” el LLM en base a feedback

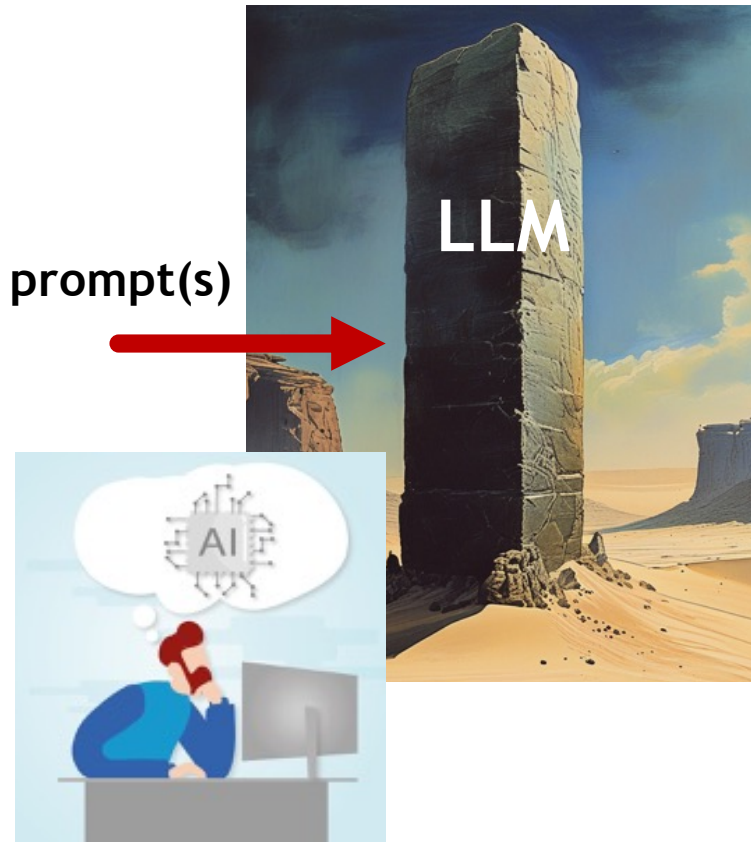


# Agregando conocimiento: RAG

- Puede verse como una alternativa al fine-tuning
- El componente de recuperación a menudo involucra una base de datos vectorial
- Provee un “esquema” para gestionar la interacción (por ej., generación) con el LLM
- El RAG (naive) puede adaptarse y complejizarse para resolver otros escenarios

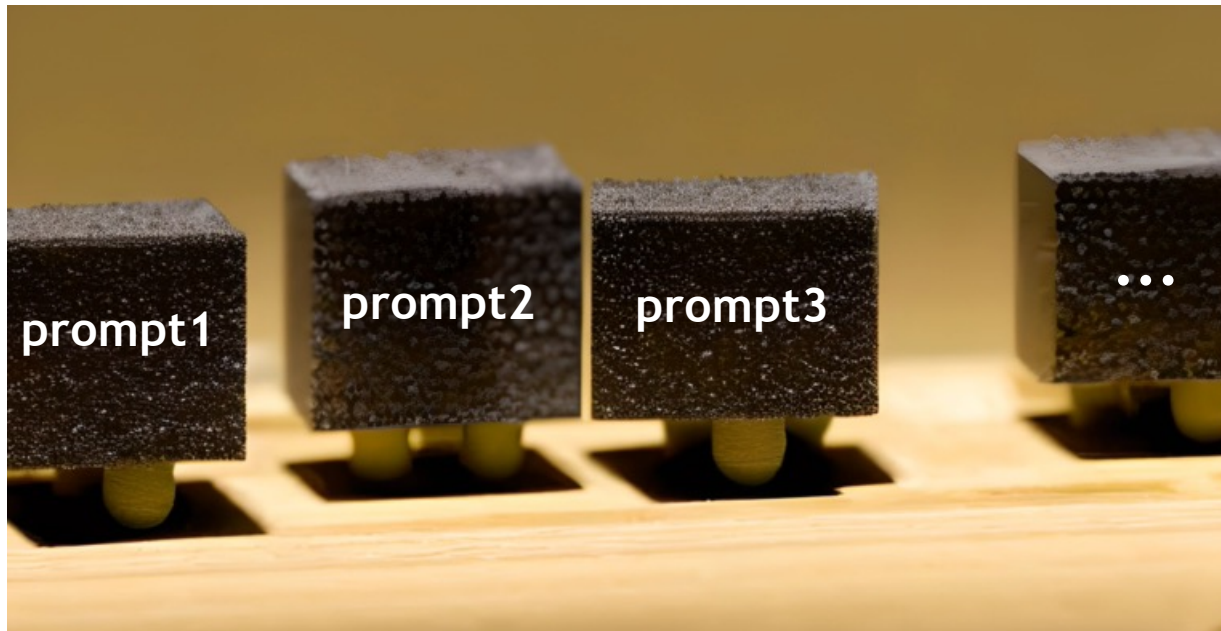


# Un punto de quiebre



- Un super-prompt no siempre es efectivo para resolver una tarea
- A medida que se desarrollan flujos con múltiples invocaciones a un LLM, el proceso constructivo se vuelve más complejo para implementar/evaluar/ajustar
- El uso de memoria y herramientas (tools) combinadas con un LLM se vuelve una necesidad

# Idea: Divide y Conquista ... + Autonomía?



# Ejemplo: VacAgent

our trip

you currently

to

country are you

n vacationing

Kenya

you are

n traveling?

24 - 03/25/2

interests and

extra details


trip?

who love

ing,

g, hiking,

ine



## VacAgent

Let AI agents plan your next vacation!

Agents at work...

Observation

Title: Eldoret Events Calendar 2024 - AllEvents.in

Link: <https://allevents.in/eldoret/calendar>

Snippet: Explore all the happening events in Eldoret in 2024 with us that best suit your interest. Theatre tickets, comedy festival, music classes or any adventure ...

Title: Conferences in Eldoret, Kenya in 2024 | Conference Locate (Clocate)

Link: <https://www.clocate.com/conferences-in+eldoret+2024/Y2ktMTE5ODcreXi=/>

Snippet: Events in Eldoret are dedicated to various topics, including science, business, technology, healthcare, social and environmental issues, lifestyle and more. We ...

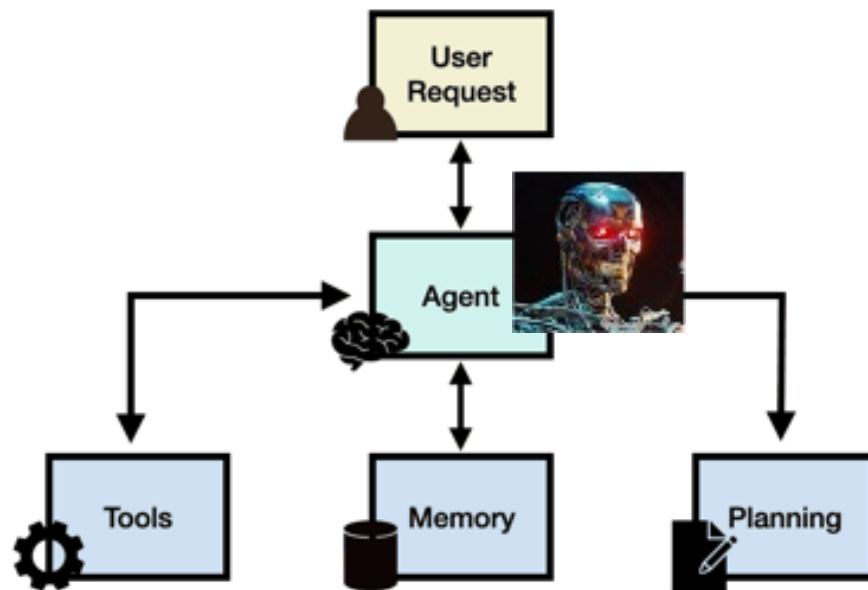


[https://github.com/tonykipkemboi/trip\\_planner\\_agent?tab=readme-ov-file](https://github.com/tonykipkemboi/trip_planner_agent?tab=readme-ov-file)



# Definición de Agente

- El concepto de agente viene desde hace mucho tiempo



- El LLM (+ prompting) actúa ahora como el cerebro del agente

- Entidad (de software) **autónoma y modular**
- Conducida por objetivos, a menudo en función de ciertas tareas
- Que se encuentra situada en un ambiente, y puede percibir señales de dicho ambiente (por ej., a través de sensores)
- Que puede ejecutar acciones sobre dicho ambiente
- Que tiene **capacidad de razonamiento**, y puede planificar “combinaciones” de acciones para lograr sus objetivos
- Puede comunicarse con otros agentes



# Memoria

- Un repositorio donde se **almacenan** distintas actividades realizadas por un agente, así como también **interacciones con el usuario**
- Distintos criterios para procesar la memoria:
  - Últimas entradas
  - Resumen
  - ...
- Corto-plazo versus largo-plazo

I'm interested in integrating LLMs with external knowledge.

LLMs are great at generating human-like text. Yet, integrating external knowledge can enhance their capabilities even more.

What are the different possible methods for doing this?

You could use pre-existing knowledge graphs, allow LLMs access to tools like APIs, or retrieval augmentation with vector DBs!

..... Conversation History .....

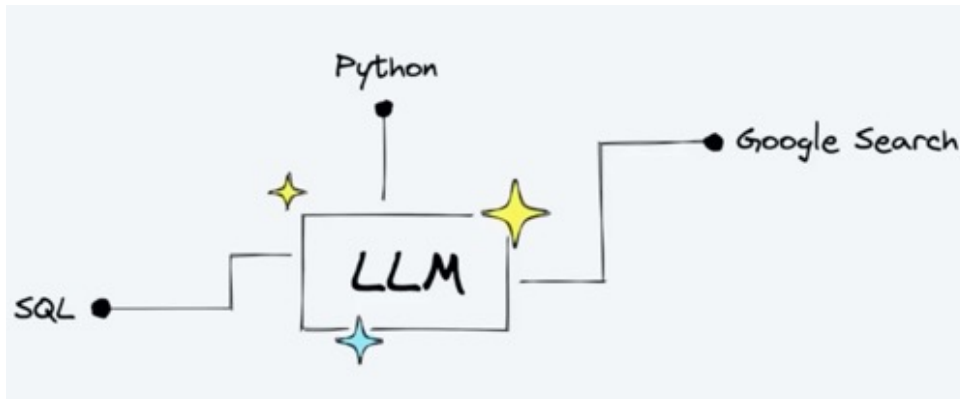
Interesting! What was it I wanted to know about again?

You were interested in integrating LLMs with external knowledge.



# Tools (function calling)

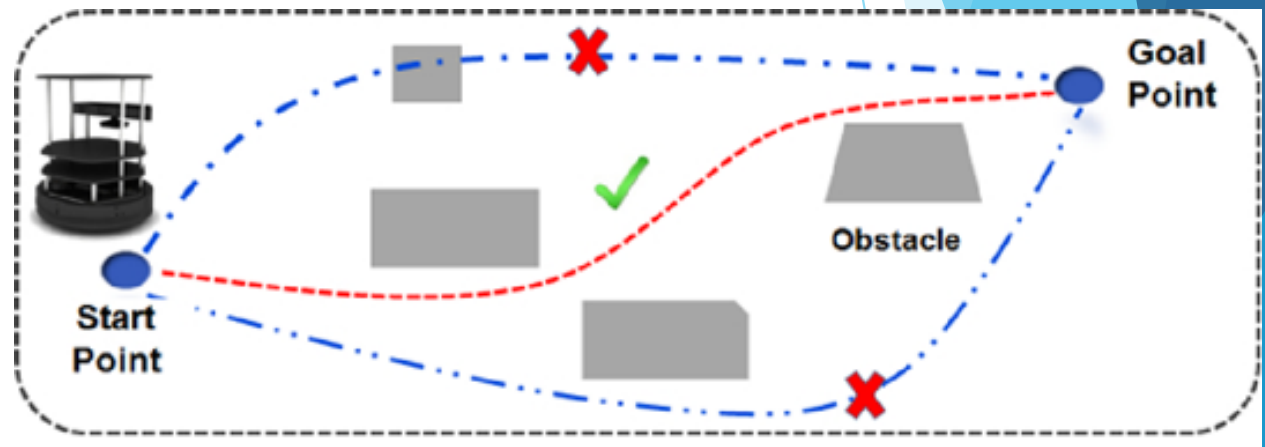
- Permiten **realizar operaciones/acciones** más allá del conocimiento del agente (o LLM) e interactuar con su ambiente
  - Acceder a información, realizar cálculos, invocar a otros sistemas, ejecutar código, etc.
  - Normalmente se expresan como funciones con parámetros de entrada y de salida
  - A menudo implementado como **function calling** en el LLM



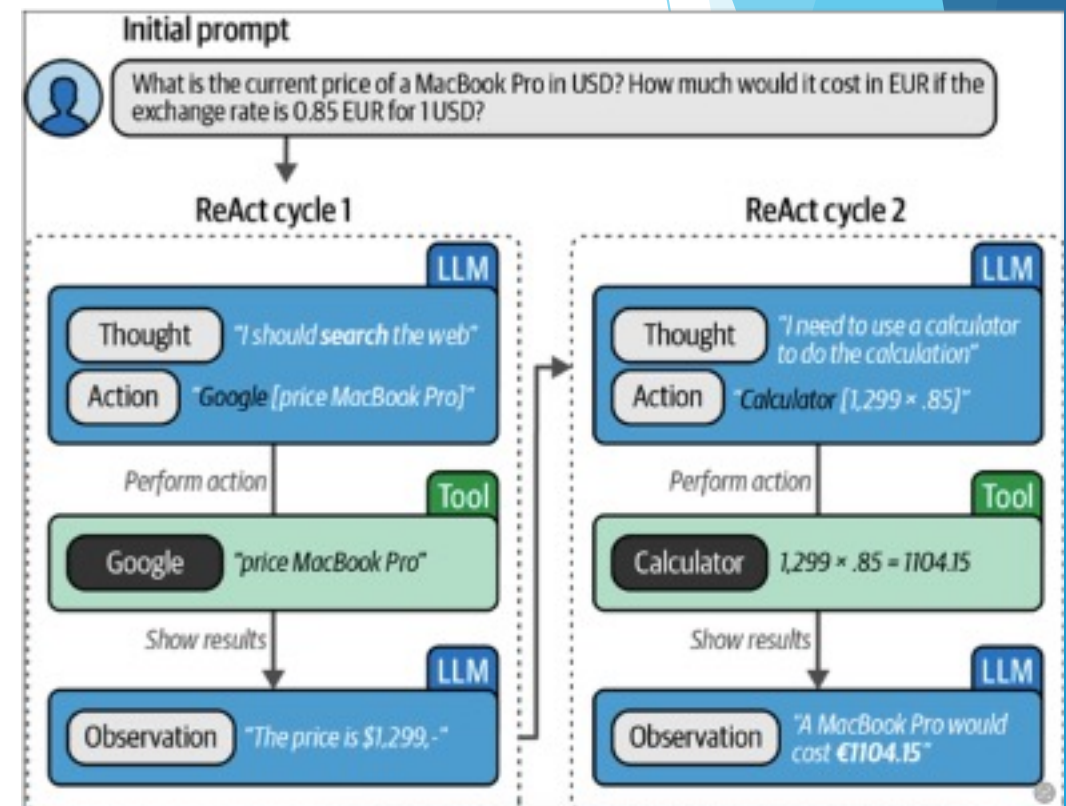
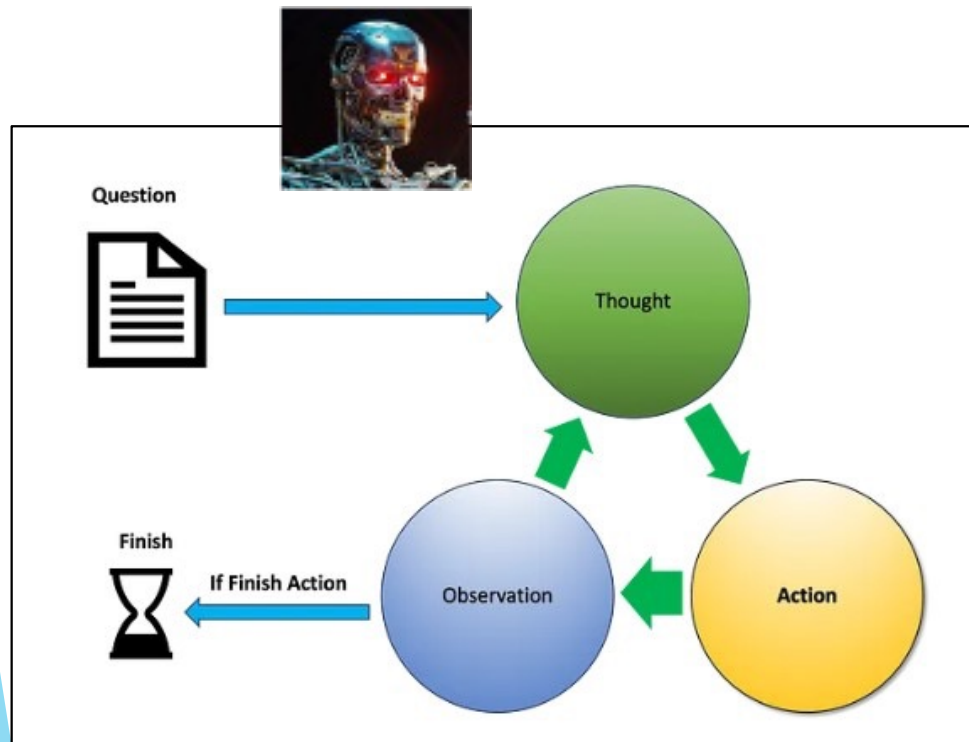
# Planificación (razonamiento)

- Abordar una pregunta (del usuario), objetivo o tarea mediante la descomposición de la misma en pasos que pueden resolverse individualmente y luego combinarse
- Distintas técnicas
  - Question decomposition
  - CoT, ToT
  - ReAct
  - Self-critic
  - ...
- Suele requerir memoria, y puede integrar tools

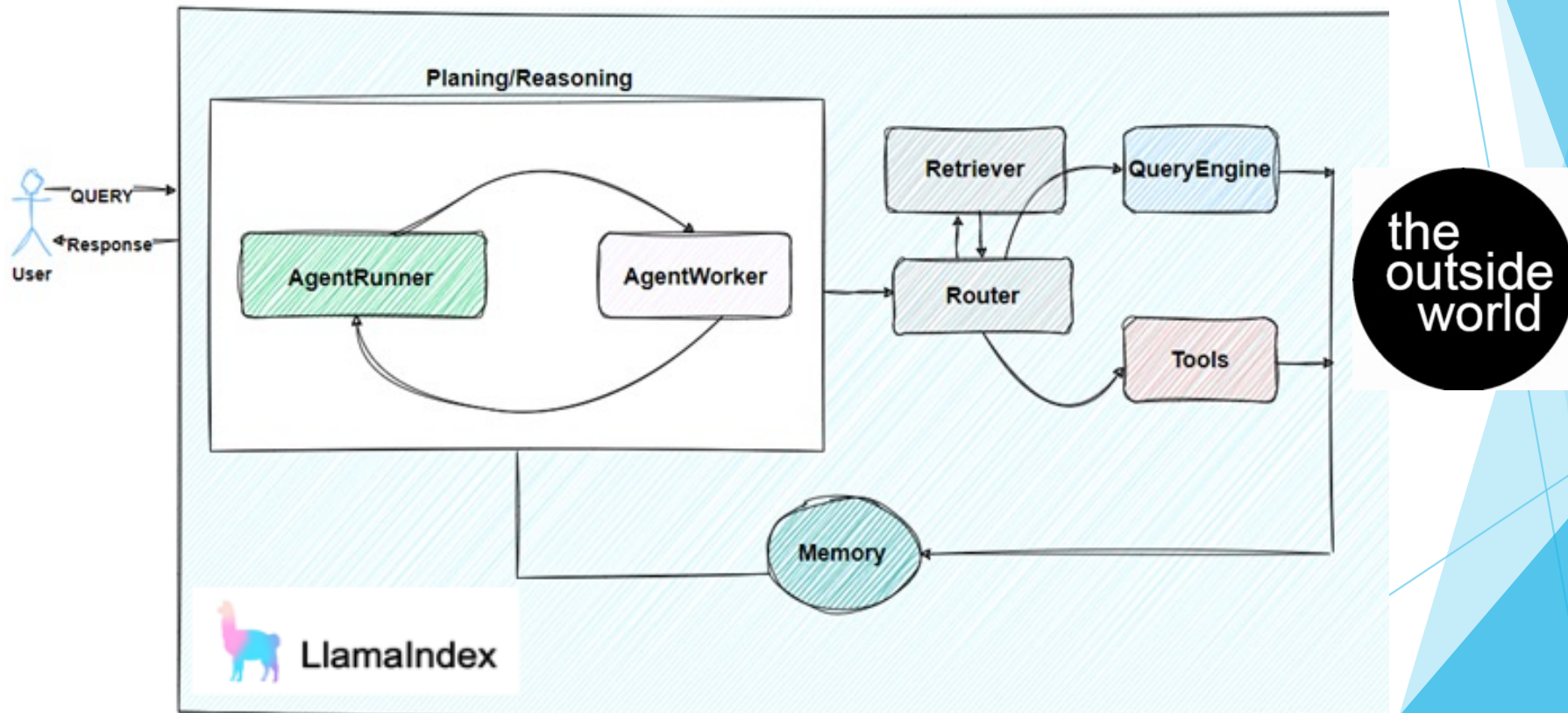
(existe un debate si LLM planning es realmente (AI) planning)



# Ejemplo: ReAct = Reasoning + Action



# Estructura interna de un agente ReAct

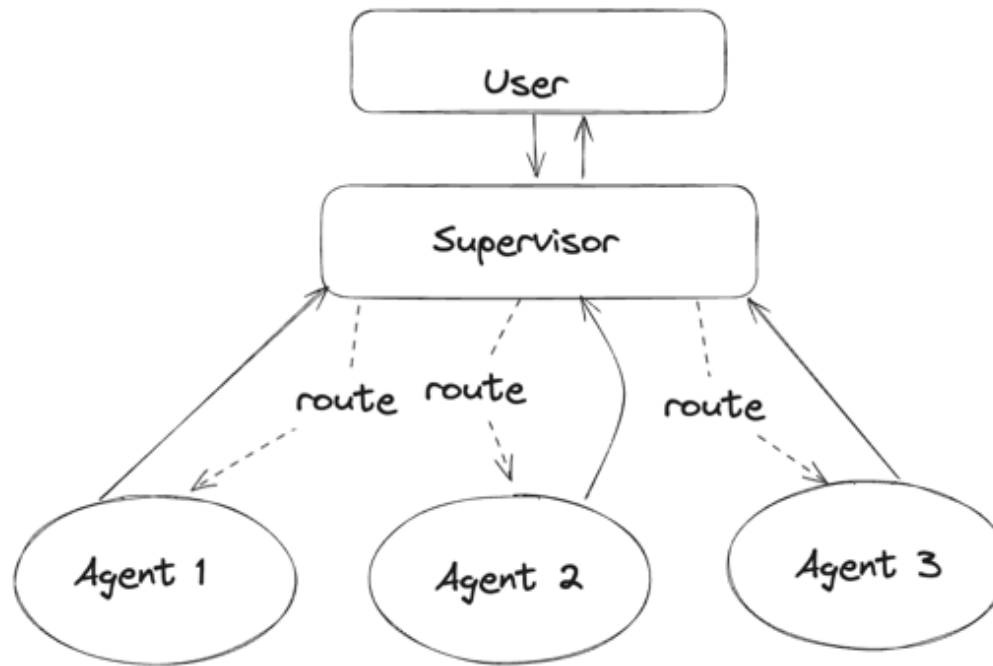




# Sistemas Multiagentes ...



# Definiendo Sistemas Multiagentes



- Un sistema que aprovecha las **capacidades de razonamiento** (potenciadas por un LLM) de distintos **agentes independientes** para tomar **decisiones** en un flujo de aplicación



# Volviendo a VacAgent



## City Selection Expert

Objetivo: Seleccionar la mejor ciudad en base a meteorología, estación, y precios



*Seleccionar ciudades en base a un punto de partida e intereses del viajero*

## Local Expert

Objetivo: Dar los puntos principales de una ciudad dada



*Compilar información detallada en forma de guía*

## Travel Concierge

Objetivo: Crear itinerarios de viaje, incluyendo sugerencias de viajes



*Expandir la guía en un itinerario, detallando un plan de actividades y lugares especiales*

## TOOLS

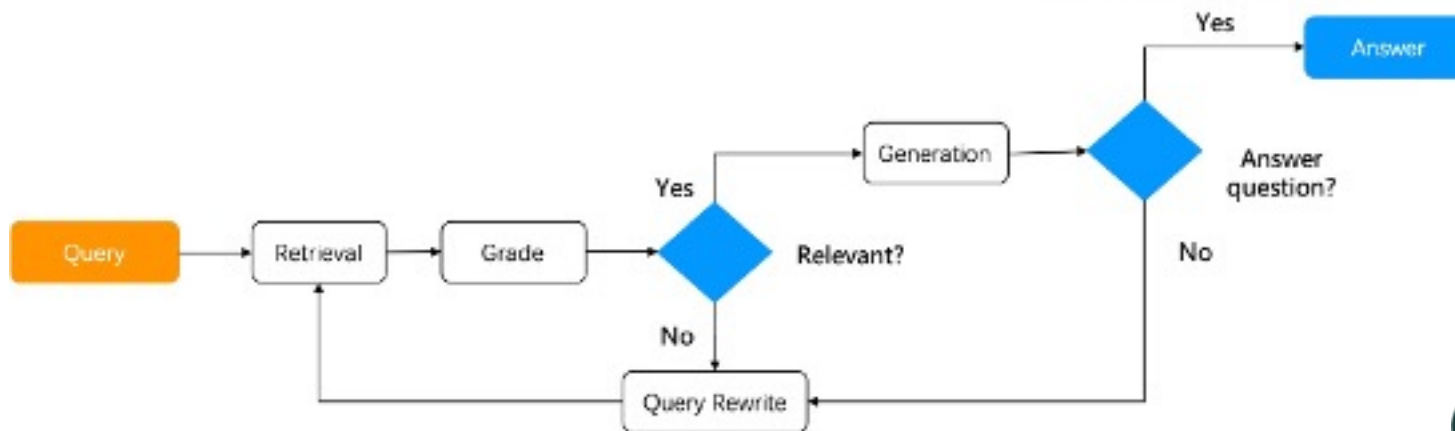
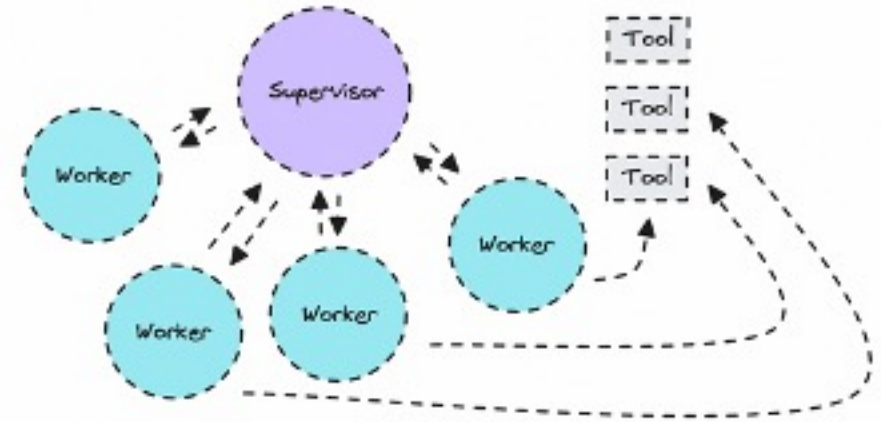


Web Scraper



# Agentic Workflows

- Descomposición en funciones, comportamiento conducido por estados
- Diseño flexible de flujos, pero reteniendo un control programático
  - Ruteo, orquestación, coreografía.

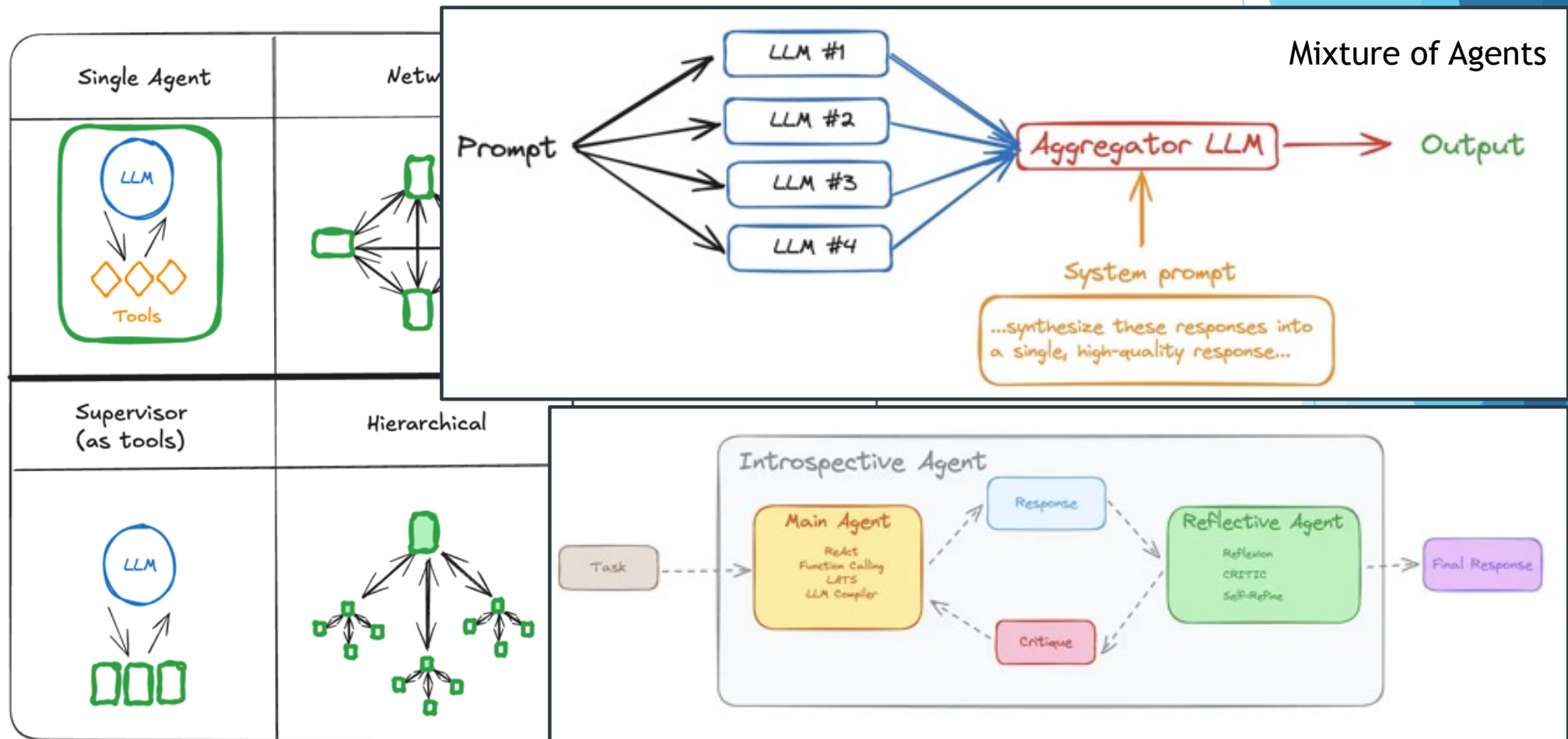


LlamaIndex



LangGraph

# Distintos patrones multi-agente



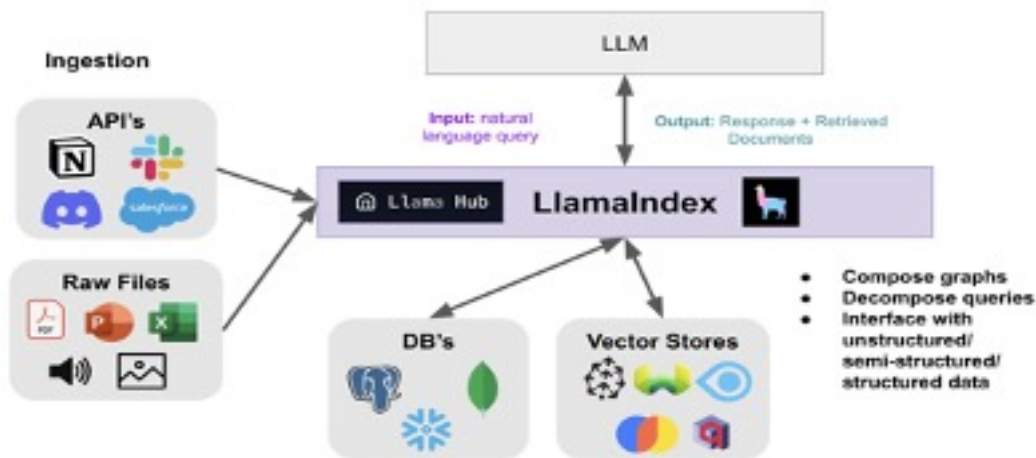
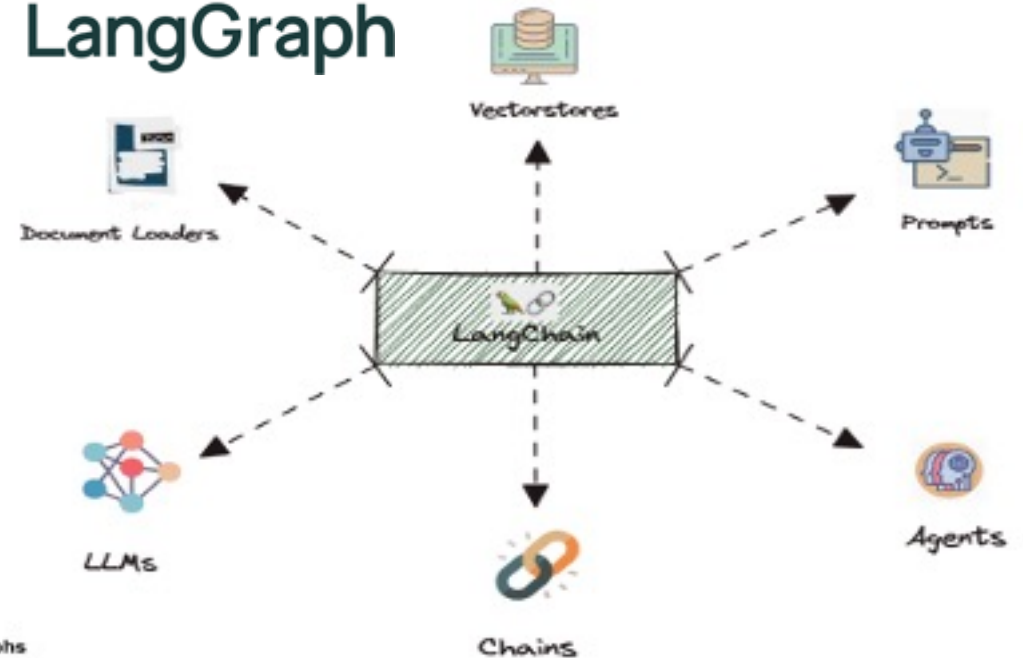
# Algunos frameworks

## microsoft/autogen

Enable Next-Gen Large Language Model Applications. Join our Discord:  
<https://discord.gg/pAbnFJrkGZ>

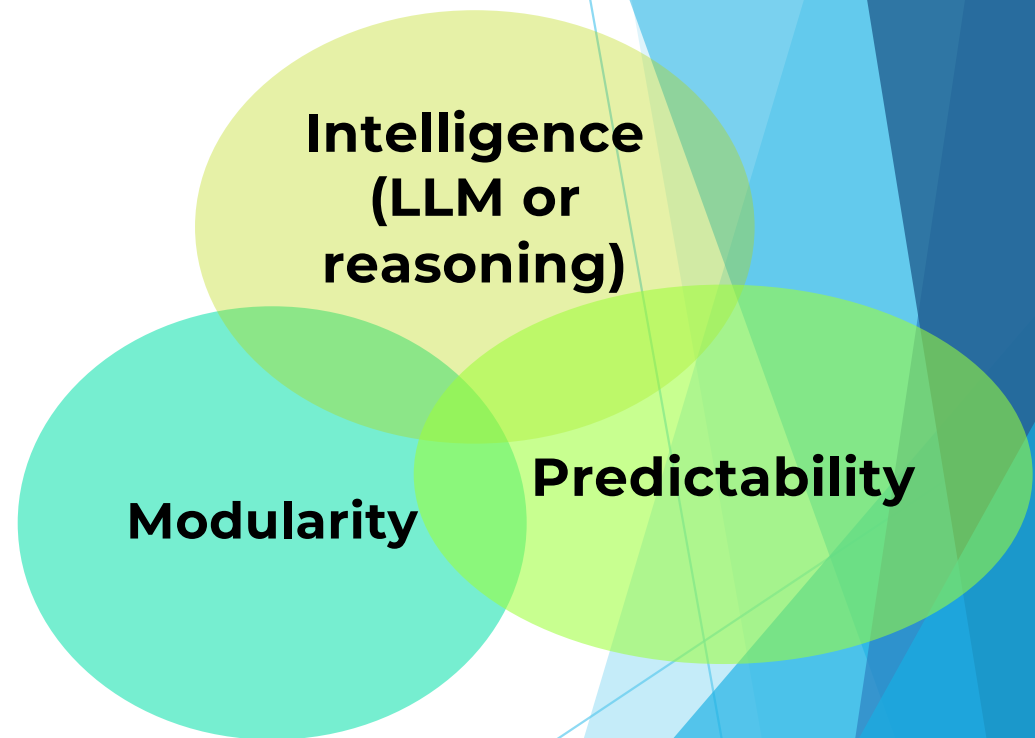


## LangGraph





















# Noción de agente: Múltiples opiniones

- Un agente es una **abstracción o patrón**
  - Funcionalidad e interfaz bien definida
  - Criterios de cohesión y acoplamiento
  - Sin embargo, la interpretación de autonomía puede tener variaciones de una aplicación a otra
- La abstracción/patrón que se necesita para una aplicación puede coincidir o no con las clases provistas por un framework/biblioteca



# Niveles de autonomía en una aplicación con LLMs

			Decide Output of Step	Decide Which Steps to Take	Decide What Steps are Available to Take
HUMAN-DRIVEN	1	Code			
	2	LLM Call	 <i>one step only</i>		
	3	Chain	 <i>multiple steps</i>		
	4	Router	 <i>no cycles</i>		
AGENT-EXECUTED	5	State Machine	 <i>cycles</i>		
	6	Autonomous			



# Lecciones aprendidas

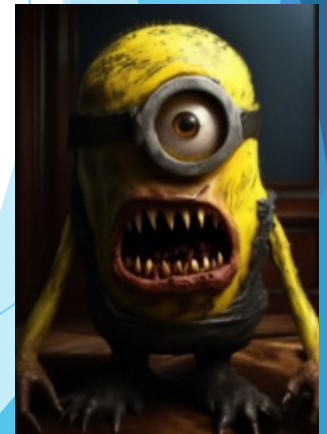


- La **modularidad** (divide y conquista) es algo beneficioso en el proceso de desarrollo
  - Los agentes/workflows pueden correrse en paralelo si hace falta
  - Los workflows basados en estado tienden a inducir acoplamientos de datos entre los pasos del mismo
- **Tradeoffs**
  - Entre mantener un control (programático) sobre los agentes (por ej., a través de un grafo) y permitir que el mismo mecanismo del agente decida qué hacer de forma autónoma
  - El agregado de pasos en un agente tiende a incrementar la latencia del flujo, pero puede mejorar la calidad de su respuesta



# Casos de Uso para Agentes

- Según Gartner, **para 2028 al menos el 15% de las decisiones de trabajo relacionadas con el día a día serán realizadas en forma autónoma por una IA con agentes**
- No solo chatbots
- Herramientas low-code y no-code
- Plugins para aplicaciones existentes
- Generadores de contenidos (por ej., para profesionales)
- Asistentes para servicio al cliente (por ej., triage)
- Personalización para self-service (por ej., consultas, reportes, analítica)
- Tutoring y educación
- Desarrollo de software
- Entretenimiento



*cool but scary*

# Desafíos

- **Testing y reproducibilidad/observabilidad** de flujos /comportamientos es importante en aplicaciones productivas
  - Asegurar que cambios en prompts (por ej., en un paso de un workflow) o en la versión del LLM no degradan comportamientos previamente chequeados es un reto
  - Se recomienda definir benchmarks (como tests de regresión), y planificar el logging de la aplicación desde el principio
- Gestión de **distintos tipos de LLMs → LLMOps** (en la misma aplicación)



# Más sobre Agentes

- Custom Agents (GPTs)

The poster is for a workshop titled 'Custom Agents' (CREACIÓN DE AGENTES DE IA INTERNOS PARA EMPRESAS). It features a dark red background with yellow and white text and graphics. At the top, there are logos for 'FLAMA' (El conocimiento prende), 'Cluster Tecnológico Tandil', and a diamond-shaped logo. The main title 'Custom Agents' is in large yellow and white font. Below it, the subtitle 'CREACIÓN DE AGENTES DE IA INTERNOS PARA EMPRESAS' is in white. A paragraph describes the workshop: 'Aprendé cómo implementar agentes de inteligencia artificial internos para optimizar procesos y mejorar servicios en tu empresa. Descubrí estrategias prácticas y accesibles para aprovechar al máximo el potencial de la IA.' Event details include: 'Salón Verde - Cámara Empresaria Mitre 856', '6 de Noviembre', '8:30 a 9:30 HS', and 'Empresas en general'. A yellow button says '¡RESERVA TU LUGAR!'. Below that is a QR code and a 'LINK' button. A circular portrait of Andrés Díaz Pace, a man with glasses and a beard, is shown. Below his name, it says 'Software Engineering'. At the bottom, there are logos for 'Cluster Tecnológico Tandil', 'CÁMARA EMPRESARIA DE TANDIL', 'CONICET', and 'Municipalidad Tandil'. The website 'WWW.FLAMA.AR' is at the bottom right.

**Custom Agents**

**CREACIÓN DE AGENTES DE IA INTERNOS PARA EMPRESAS**

Aprendé cómo implementar agentes de inteligencia artificial internos para optimizar procesos y mejorar servicios en tu empresa. Descubrí estrategias prácticas y accesibles para aprovechar al máximo el potencial de la IA.

Salón Verde - Cámara Empresaria Mitre 856  
6 de Noviembre 8:30 a 9:30 HS  
Empresas en general

**¡RESERVA TU LUGAR!**

**ANDRÉS DIAZ PACE**  
Software Engineering

**LINK**

Cluster Tecnológico Tandil  
CÁMARA EMPRESARIA DE TANDIL  
CONICET  
Municipalidad Tandil

**WWW.FLAMA.AR**

**Gracias!**

**Y los  
esperamos  
mañana!**

**Andrés Díaz Pace**

**1 & 2 / Nov / 2024**

[andres.diazpace@isistan.unicen.edu.ar](mailto:andres.diazpace@isistan.unicen.edu.ar)

[andres.diazpace@globant.com](mailto:andres.diazpace@globant.com)



**{GEERS}**

