

MANUAL

GUÍA DE REFERENCIA COMPLETA



# Cómo funciona un agente de IA

*Arquitectura, funcionamiento interno, capacidades, límites y buenas prácticas*

CLASE 2

DOCENTE: EMILSE KASEM

# AGENDA

1 · Qué es un agente de IA

2 · La arquitectura de un agente: sus piezas internas

3 · El ciclo de funcionamiento: cómo piensa y actúa un agente

4 · Tipos de agentes y niveles de autonomía

5 · Memoria y aprendizaje: qué recuerda un agente y qué no

6 · Casos de uso aplicados

7 · Limitaciones técnicas

8 · Riesgos y consideraciones de seguridad

9 · Buenas prácticas para trabajar con agentes de IA



# QUÉ ES UN AGENTE DE IA

Un agente de IA es un sistema de software que usa un modelo de lenguaje de gran escala (LLM) como motor de razonamiento para perseguir un objetivo de forma relativamente autónoma: percibe información de su entorno, decide qué acción tomar, ejecuta esa acción usando herramientas reales y observa el resultado para decidir el siguiente paso.

Ese ciclo de percibir, decidir, actuar y observar puede repetirse muchas veces hasta que el objetivo se considera cumplido.



## LA PALABRA CLAVE

### Autonomía relativa

No solo responde una pregunta con texto, como un chatbot simple: puede tomar iniciativa dentro de límites definidos, encadenar pasos sin que un humano indique cada uno, y ajustar su plan si algo no sale como esperaba.

## CLASE 2

DOCENTE: **EMILSE KASEM**

# MODELO DE IA VS. AGENTE DE IA



*La analogía más clara es la de un motor y un automóvil: el motor genera potencia, pero por sí solo no te lleva a ningún lado; necesita carrocería, volante, ruedas y navegación para convertirse en un medio de transporte utilizable.*

## EL MODELO DE IA — el motor

Red neuronal entrenada que, dado un texto de entrada, predice el texto de salida más probable. Por sí solo no tiene memoria persistente entre conversaciones, no puede ejecutar acciones en el mundo real, no sabe qué pasó después de su entrenamiento y solo responde a lo que se le pregunta, sin iniciativa propia.

## EL AGENTE DE IA — el auto completo

Arquitectura de software que envuelve al modelo y le agrega lo que le falta: un objetivo definido, acceso a herramientas para actuar sobre sistemas reales, memoria de corto y largo plazo, un ciclo de razonamiento para planificar varios pasos, y mecanismos de control y supervisión.

CLASE 2

DOCENTE: **EMILSE KASEM**

# LOS CUATRO INGREDIENTES MÍNIMOS



Más allá de la complejidad de cada implementación, todo agente de IA —desde el más simple hasta el más sofisticado— comparte estos cuatro ingredientes:

## 1 Un objetivo

Una meta o tarea a cumplir, definida por quien diseña el agente o por quien lo usa en el momento.

## 2 Percepción del entorno

La capacidad de recibir información relevante: un mensaje, un archivo, el resultado de una búsqueda, una notificación.

## 3 Capacidad de decidir y actuar

Un motor de razonamiento que interpreta la situación y decide qué hacer, y herramientas reales para ejecutarlo.

## 4 Un ciclo de retroalimentación

La capacidad de observar el resultado de sus propias acciones y ajustar el siguiente paso en consecuencia.

**CLASE 2**

**DOCENTE: EMILSE KASEM**

# LA ARQUITECTURA DE UN AGENTE

Si se abre la "caja" de un agente de IA moderno, se encuentran siete componentes que trabajan de forma coordinada. No todos los agentes los implementan con la misma sofisticación, pero conceptualmente todos están presentes en cualquier sistema agéntico serio.



- 1 Definición del objetivo
- 2 Percepción y procesamiento de entradas
- 3 Memoria
- 4 Motor de razonamiento y planificación
- 5 Ejecución de herramientas (tool use)
- 6 Orquestación y coordinación
- 7 Retroalimentación y observabilidad

# OBJETIVO Y PERCEPCIÓN



## 2.1 · Definición del objetivo

Es la capa que ancla el propósito del agente: establece qué está tratando de lograr, cómo se mide el éxito, qué límites no puede cruzar (restricciones de negocio, de compliance, de seguridad) y en qué condición debe detenerse.

Un agente sin un objetivo bien definido tiende a divagar o actuar fuera del alcance esperado. Definir bien el objetivo es, en la práctica, la parte más importante del diseño de un agente.

## 2.2 · Percepción y procesamiento

Es el "sistema sensorial" del agente: la capa que recibe información cruda —texto, un documento, la respuesta de una API, una imagen— y la transforma en un formato que el motor de razonamiento pueda interpretar.

También valida y filtra la entrada: descarta información irrelevante y detecta contenido malicioso, dado que la ventana de contexto del modelo es limitada.

**CLASE 2**

**DOCENTE: EMILSE KASEM**

# MEMORIA



Es uno de los componentes más determinantes en qué tan útil resulta un agente en tareas largas o recurrentes. En la práctica se distinguen varios tipos, que suelen combinarse:

## Memoria de trabajo (corto plazo)

Es el contexto inmediato de la conversación o tarea en curso. Desaparece al terminar la sesión, salvo que se guarde explícitamente.

## Memoria persistente (largo plazo)

Información guardada de forma duradera y recuperable en sesiones futuras: preferencias, datos de proyectos anteriores, decisiones tomadas.

## Memoria episódica

Registra eventos concretos con su contexto temporal —qué se hizo, cuándo y por qué—, clave para auditar después las decisiones del agente.

## Memoria semántica / RAG

Permite al agente "buscar" información relevante en una base de conocimiento externa en el momento en que la necesita.

**CLASE 2**

**DOCENTE: EMILSE KASEM**

# MOTOR DE RAZONAMIENTO Y PLANIFICACIÓN



Es el "cerebro" del agente: el modelo de lenguaje que, a partir del objetivo, la información percibida y lo que recuerda, decide qué hacer a continuación. Existen distintos patrones de razonamiento:

## Razonar y actuar (ReAct)

Alternan ciclos cortos de observar → razonar → actuar. El más flexible: se adapta sobre la marcha, aunque consume más recursos.

## Planificar y ejecutar (Plan-and-Execute)

Arma primero un plan completo y después lo ejecuta de punta a punta. Funciona bien en tareas estables.

## Cadena de razonamiento (Chain-of-Thought)

El modelo "piensa en voz alta", paso a paso, antes de decidir, mejorando la calidad en tareas de varios pasos lógicos.

## Árbol de alternativas (Tree-of-Thought)

Explora varias líneas de razonamiento en paralelo antes de elegir la mejor, útil para comparar caminos posibles.

**CLASE 2**

**DOCENTE: EMILSE KASEM**

# HERRAMIENTAS, ORQUESTACIÓN Y RETROALIMENTACIÓN



## 2.5 · Ejecución de herramientas

Un agente puede invocar herramientas reales —buscadores, archivos, calendarios, APIs— para producir un efecto en el mundo. El estándar MCP (Model Context Protocol) permite conectarse a cualquier herramienta de forma común, como un puerto USB.

## 2.6 · Orquestación y coordinación

En tareas complejas, un agente "supervisor" descompone el trabajo en subtarefas y las distribuye entre agentes "trabajadores" especializados, integrando después los resultados.

## 2.7 · Retroalimentación y observabilidad

Registra qué decidió el agente, por qué, qué herramientas usó y qué resultado obtuvo. Es condición para confiar en el agente en contextos donde las decisiones importan.

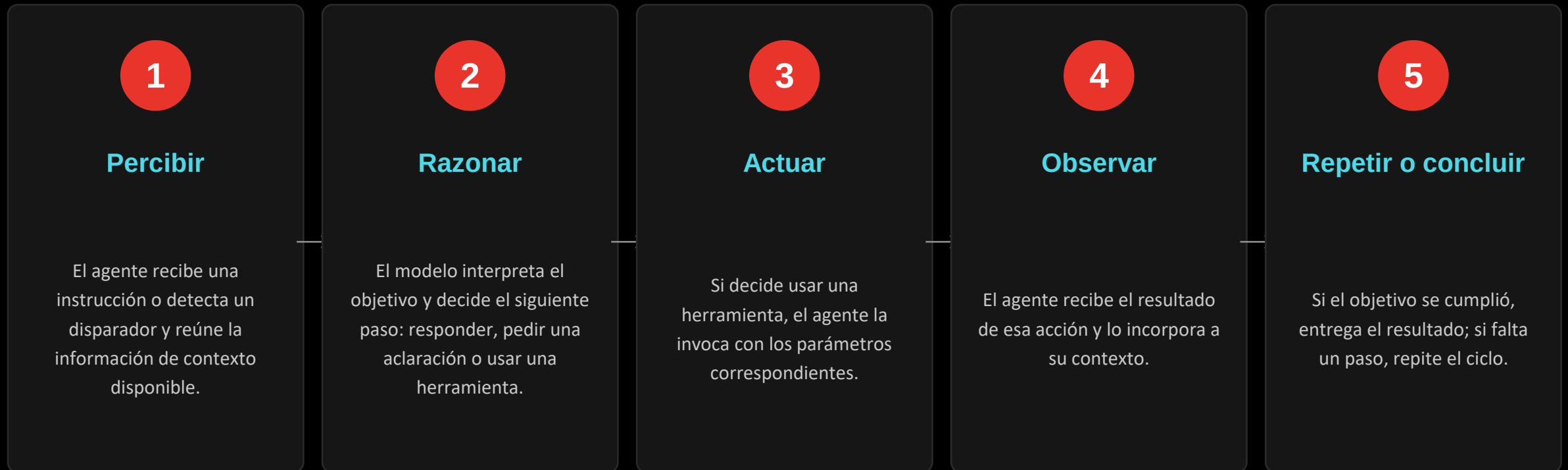
**CLASE 2**

**DOCENTE: EMILSE KASEM**

# EL CICLO DE FUNCIONAMIENTO



*El ciclo que se repite en cada paso: percibir, razonar, actuar y observar.*



**CLASE 2**

**DOCENTE: EMILSE KASEM**

# COORDINAR UNA REUNIÓN CON UN PROVEEDOR



*Instrucción: "Coordiná una reunión de una hora con el proveedor de materiales para la semana que viene, en un horario que me quede libre, y mandale la propuesta de precios actualizada."*

| PASO                | QUÉ HACE EL AGENTE                                                                                                      |
|---------------------|-------------------------------------------------------------------------------------------------------------------------|
| 1 • Percibir        | Identifica el objetivo —agendar una reunión y enviar un archivo— y las restricciones: duración, semana, disponibilidad. |
| 2 • Razonar         | Decide que primero necesita consultar el calendario para ver qué horarios están libres.                                 |
| 3 • Actuar          | Invoca la herramienta de calendario y consulta la disponibilidad de la semana próxima.                                  |
| 4 • Observar        | Recibe tres horarios libres y, como el objetivo también incluye enviar la propuesta, ubica ese archivo.                 |
| 5 • Actuar de nuevo | Busca en los archivos recientes la "propuesta de precios actualizada" y la encuentra.                                   |
| 6 • Actuar          | Redacta la invitación con un horario disponible, adjunta el archivo y la envía al contacto del proveedor.               |
| 7 • Concluir        | Informa que la reunión quedó coordinada y que la propuesta fue enviada — o se detiene a preguntar si hubo ambigüedad.   |

CLASE 2

# TIPOS DE AGENTES SEGÚN SU FORMA DE DECIDIR



| TIPO                               | CÓMO DECIDE                                                                                                 |
|------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Reactivos simples                  | Responden a una entrada con una acción predefinida, sin planificación ni memoria de lo anterior.            |
| Deliberativos (basados en modelos) | Mantienen una representación interna de su entorno para anticipar el efecto de sus acciones.                |
| Basados en objetivos               | Evalúan distintas secuencias posibles de acciones y eligen la que mejor los acerca a un objetivo explícito. |
| Basados en utilidad                | Comparan resultados posibles según una medida de "qué tan bueno" es cada uno.                               |
| De aprendizaje                     | Ajustan su comportamiento con el tiempo a partir de la retroalimentación recibida.                          |
| Sistemas multiagente               | Varios agentes especializados colaboran, coordinados por una capa de orquestación.                          |

CLASE 2

DOCENTE: EMILSE KASEM

# NIVELES DE AUTONOMÍA



## Asistido

El agente sugiere una acción, pero una persona la revisa y la ejecuta manualmente. El nivel más seguro para empezar.

## Semi-autónomo (con aprobación)

El agente prepara y deja lista la acción, pero espera una confirmación explícita antes de ejecutarla.

## Autónomo con supervisión

El agente ejecuta la acción por su cuenta, pero queda un registro auditable y puntos de control para revisar o revertir.

## Autónomo pleno

Decide y ejecuta sin intervención humana en el momento. Mayor productividad, pero también mayor riesgo.

*La recomendación general es no asignar autonomía plena desde el primer día, sino escalar progresivamente a medida que se verifica que las decisiones del agente son confiables.*

**CLASE 2**

DOCENTE: **EMILSE KASEM**

# EL AGENTE NO "APRENDE" COMO UNA PERSONA



Uno de los malentendidos más comunes es pensar que un agente de IA aprende de cada conversación de la misma forma en que aprende una persona con la experiencia. En realidad, el modelo de lenguaje que actúa como motor de razonamiento queda fijo una vez entrenado: no modifica sus "conocimientos" internos por haber tenido una conversación.

## ¿QUÉ ES ENTONCES LA SENSACIÓN DE "APRENDIZAJE"?

Que el agente guarde información en su memoria persistente (una base de datos externa al modelo) y la recupere en conversaciones futuras. Eso no es el modelo aprendiendo, sino el sistema de memoria guardando y recuperando datos —de forma parecida a como una persona usa una libreta de notas en lugar de memorizarlo todo.

Por eso: un agente puede "olvidar" algo si no quedó guardado explícitamente, y corregirlo en una conversación no evita que repita el error con otro usuario, salvo que la corrección se guarde como regla.

## CLASE 2

DOCENTE: **EMILSE KASEM**

# CASOS DE USO APLICADOS

## Gestión administrativa y de oficina

Leer el correo entrante, redactar borradores, actualizar una planilla de seguimiento y recordar vencimientos, dejando la revisión final a la persona.

## Atención al cliente

Conectado al historial de pedidos y a la base de políticas, resolver consultas frecuentes y derivar los casos que superan su nivel de confianza.

## Investigación y síntesis

Buscar en múltiples fuentes, comparar información y producir un resumen estructurado con citas a las fuentes originales.

## Redes sociales y contenido

Proponer un calendario de publicaciones, redactar borradores en el tono de la marca y programar su publicación, con aprobación final humana.

## Gestión de proyectos

Monitorear el avance de tareas contra un cronograma, detectar desvíos y armar minutas de reunión a partir de notas dispersas.



# LIMITACIONES TÉCNICAS



| LIMITACIÓN                            | EN QUÉ CONSISTE                                                                                                               |
|---------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| Alucinaciones                         | Puede producir información que suena coherente pero es incorrecta o inventada, sobre todo con datos muy específicos.          |
| Ventana de contexto finita            | Solo puede "tener en mente" una cantidad limitada de información a la vez; puede decidir con una imagen parcial.              |
| Sin comprensión ni lógica garantizada | Predice la continuación más probable de un texto; no hay garantía matemática de corrección.                                   |
| Dependencia de datos y prompt         | Instrucciones ambiguas o datos de baja calidad se traducen en decisiones y resultados de baja calidad.                        |
| Fallo compuesto en varios pasos       | Cada paso tiene una probabilidad de error; en tareas largas, los errores se acumulan de forma multiplicativa.                 |
| Costo, latencia y variabilidad        | Cada paso de razonamiento tiene un costo y un tiempo asociado, y el mismo agente puede variar su respuesta entre ejecuciones. |

CLASE 2

DOCENTE: EMILSE KASEM

# RIESGOS Y CONSIDERACIONES DE SEGURIDAD



| RIESGO                             | EN QUÉ CONSISTE                                                                                                                  |
|------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|
| Inyección de instrucciones         | Contenido externo (una página, un correo) puede incluir instrucciones ocultas que manipulan al agente.                           |
| Filtración de información sensible | Un agente con acceso a múltiples fuentes puede combinar información de contextos distintos de forma indebida.                    |
| Sesgos heredados                   | Puede reproducir sesgos sociales o culturales presentes en los datos con los que fue entrenado el modelo.                        |
| Autonomía sin supervisión adecuada | A mayor autonomía, mayor daño potencial de una decisión equivocada, en particular si es difícil de revertir.                     |
| Falta de trazabilidad              | Un agente que no deja registro de qué decidió y por qué es, en la práctica, imposible de auditar.                                |
| Responsabilidad y gobernanza       | Quién responde por una decisión del agente sigue siendo una cuestión abierta: tratarlo como herramienta bajo supervisión humana. |

# BUENAS PRÁCTICAS PARA TRABAJAR CON AGENTES DE IA



1

Definir el objetivo y sus límites con precisión.

2

Empezar con autonomía baja y escalar gradualmente.

3

Reservar la aprobación humana para acciones irreversibles o de alto impacto.

4

Verificar los resultados críticos antes de darlos por buenos.

5

Mantener registros y auditoría de cada decisión.

6

Limitar el acceso a herramientas según el principio de necesidad mínima.

7

Tratar el contenido externo como información, nunca como instrucciones.

8

Revisar y actualizar el agente periódicamente.

CLASE 2

DOCENTE: **EMILSE KASEM**

# LO ESENCIAL DE ESTE MANUAL



Un agente de IA no es magia ni es, en el otro extremo, "solo un chatbot con otro nombre": es una arquitectura de software concreta que envuelve a un modelo de lenguaje con objetivo, percepción, memoria, capacidad de usar herramientas reales y mecanismos de control, para poder actuar de forma relativamente autónoma sobre un problema.

## EL PRINCIPIO QUE ATRAVIESA TODO EL MANUAL

**Cuanto mayor el impacto o la irreversibilidad de una decisión, mayor debe ser el nivel de supervisión humana sobre el agente que la toma.**

*Con ese principio como guía, los agentes de IA son, hoy, una de las herramientas más útiles disponibles para multiplicar la capacidad de trabajo de una persona o de un equipo.*

## CLASE 2

DOCENTE: **EMILSE KASEM**

# GRACIAS

Cómo funciona un agente de IA

*Arquitectura, funcionamiento interno, capacidades, límites y buenas prácticas*