



Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Arquitectura de memoria para agentes de inteligencia artificial

Cómo elegir el tipo correcto según el caso de uso

Cuando se diseña un agente de inteligencia artificial, una de las decisiones más importantes es **cómo manejar la memoria del sistema**.

Muchos proyectos comienzan utilizando una única fuente de datos, como una hoja de cálculo o una base de datos simple. Sin embargo, los sistemas reales de agentes suelen requerir **varios tipos de memoria trabajando en conjunto**.

Cada tipo de memoria resuelve un problema distinto:

- almacenar datos exactos
- mantener el contexto de conversación
- buscar información en documentos
- recordar quién es el usuario

Este documento explica **qué tipo de memoria utilizar según el caso de uso** y cómo combinarlas en una arquitectura coherente.

Los cuatro tipos de memoria en agentes de IA

Antes de ver los casos de uso, repasemos brevemente los cuatro tipos principales.



Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Memoria estructurada

Se utiliza para **datos exactos y verificables**.

Ejemplos:

- inventario
- turnos disponibles
- datos de clientes
- estados de pago
- registros en un CRM

Esta memoria suele almacenarse en:

- bases de datos
- APIs externas
- sistemas internos
- hojas de cálculo

El agente debe **consultar esta información**, no generarla.

Memoria conversacional

Permite mantener **el contexto de la conversación**.

Sin esta memoria, el agente no puede entender referencias a mensajes anteriores.

Ejemplo:

Usuario:

“Necesito un turno para la semana que viene.”

Luego dice:

“Mejor el martes.”



Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

El sistema debe entender que el martes se refiere al turno mencionado previamente.

Para este tipo de memoria generalmente se puede usar MySQL o Redis. Personalmente recomiendo Redis ya que en situaciones normales un usuario puede enviar más de un mensaje en un lapso de tiempo corto y se solaparía con las respuestas que va generando el agente.

Memoria semántica

Permite **buscar conocimiento en documentos o bases de información**.

Se implementa generalmente con:

- embeddings
- bases de datos vectoriales
- sistemas RAG

Esto permite recuperar información incluso cuando el usuario utiliza palabras distintas a las del documento original.

Memoria basada en identidad

Permite asociar información a **un usuario específico**.

Se puede vincular a:

- número de teléfono
- cuenta de WhatsApp
- correo electrónico
- ID de usuario

Esto permite que el sistema recuerde el historial de interacciones, preferencias o datos previos. Para este tipo de memoria generalmente utilizamos MySQL, Postgres, u otro motor de bases de datos relacional.



Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Casos de uso y arquitectura de memoria recomendada

A continuación se presentan algunos casos de uso comunes en agentes de IA y las memorias más adecuadas para cada uno.

Agentes de atención al cliente

Ejemplo:

- responder preguntas frecuentes
- consultar información sobre productos o servicios
- guiar al usuario en procesos

Memorias recomendadas:

Memoria semántica

Para buscar respuestas en documentación o FAQs.

Memoria conversacional

Para mantener el contexto del diálogo.

Memoria basada en identidad

Para recordar conversaciones previas del cliente.

Memoria estructurada (opcional)

Para consultar datos específicos como pedidos o cuentas.



Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Agentes para agendar citas

Ejemplo:

- clínicas
- peluquerías
- servicios profesionales

Memorias recomendadas:

Memoria estructurada

Para consultar turnos disponibles y crear reservas.

Memoria conversacional

Para mantener el contexto del proceso de agendamiento.

Memoria basada en identidad

Para recordar citas anteriores o datos del cliente.

Memoria semántica (opcional)

Para responder preguntas sobre servicios o procedimientos.

Agentes de soporte técnico

Ejemplo:

- resolver problemas de software
- explicar configuraciones
- asistir en procesos técnicos

Memorias recomendadas:

Memoria semántica

Para buscar información dentro de documentación técnica.



Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Memoria conversacional

Para seguir el flujo del diagnóstico.

Memoria basada en identidad

Para conocer el historial del cliente o del problema.

Memoria estructurada (opcional)

Para consultar tickets o estados de incidencias.

Agentes de ventas o generación de leads

Ejemplo:

- responder preguntas sobre productos
- calificar potenciales clientes
- guiar el proceso de compra

Memorias recomendadas:

Memoria conversacional

Para mantener el flujo de la conversación.

Memoria basada en identidad

Para registrar información del lead.

Memoria semántica

Para responder preguntas sobre productos o servicios.

Memoria estructurada

Para guardar datos del cliente o del lead.



Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Agentes internos para empresas

Ejemplo:

- consultas sobre procesos internos
- búsqueda de información en documentos corporativos
- asistencia a empleados

Memorias recomendadas:

Memoria semántica

Para buscar información en documentación interna.

Memoria conversacional

Para mantener contexto en consultas complejas.

Memoria basada en identidad

Para personalizar respuestas según el usuario.

Memoria estructurada (opcional)

Para consultar sistemas internos de la empresa.

Errores comunes al diseñar memoria para agentes

Usar una sola memoria para todo

Un error muy común es intentar resolver todo con una única fuente de datos.

Cada tipo de memoria resuelve un problema distinto.



Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Permitir que el agente invente datos estructurados

Datos como inventarios o turnos **siempre deben consultarse en sistemas reales**.

Nunca deberían ser generados por el modelo.

No separar conocimiento de datos operativos

Los documentos y bases de conocimiento deberían manejarse con **memoria semántica**, no con bases de datos tradicionales.

No vincular las conversaciones a un usuario

Sin memoria basada en identidad, el sistema no puede ofrecer experiencias personalizadas.

Regla simple para diseñar agentes

Antes de elegir herramientas o tecnologías, define:

- qué tipo de información necesita el agente
- qué tipo de memoria corresponde a esa información

Una vez definida la arquitectura de memoria, recién entonces tiene sentido elegir herramientas o frameworks.

Primero arquitectura.
Después tecnología.



Tabla de selección de memoria

Esta tabla sirve como una referencia rápida para decidir **qué tipo de memoria utilizar según el tipo de información**.

Tipo de información	Tipo de memoria recomendada	Ejemplo	Tecnología común
Inventario o disponibilidad	Memoria estructurada	Turnos de una clínica	Base de datos, API, CRM
Datos de clientes	Memoria estructurada	Historial de pedidos	CRM, base de datos
Contexto de conversación	Memoria conversacional	Usuario cambia un turno	Historial de mensajes
Documentación o FAQs	Memoria semántica	Manuales de producto	Vector DB, embeddings
Base de conocimiento	Memoria semántica	Procedimientos internos	RAG
Historial del usuario	Memoria basada en identidad	Conversaciones previas	User ID, teléfono
Preferencias del cliente	Memoria basada en identidad	Idioma preferido	Perfil de usuario
Diagnóstico en soporte	Conversacional + semántica	Troubleshooting	Chat history + RAG
Reservas o pagos	Memoria estructurada	Crear cita	API / base de datos
Personalización	Identidad + estructurada	Cliente VIP	CRM + user profile



Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Sobre este documento

Este documento forma parte de una serie de recursos sobre **arquitectura de agentes de inteligencia artificial**.

Si te interesa aprender cómo diseñar sistemas de agentes para **WhatsApp, voz o atención al cliente**, puedes encontrar más contenido técnico en mi canal de YouTube.