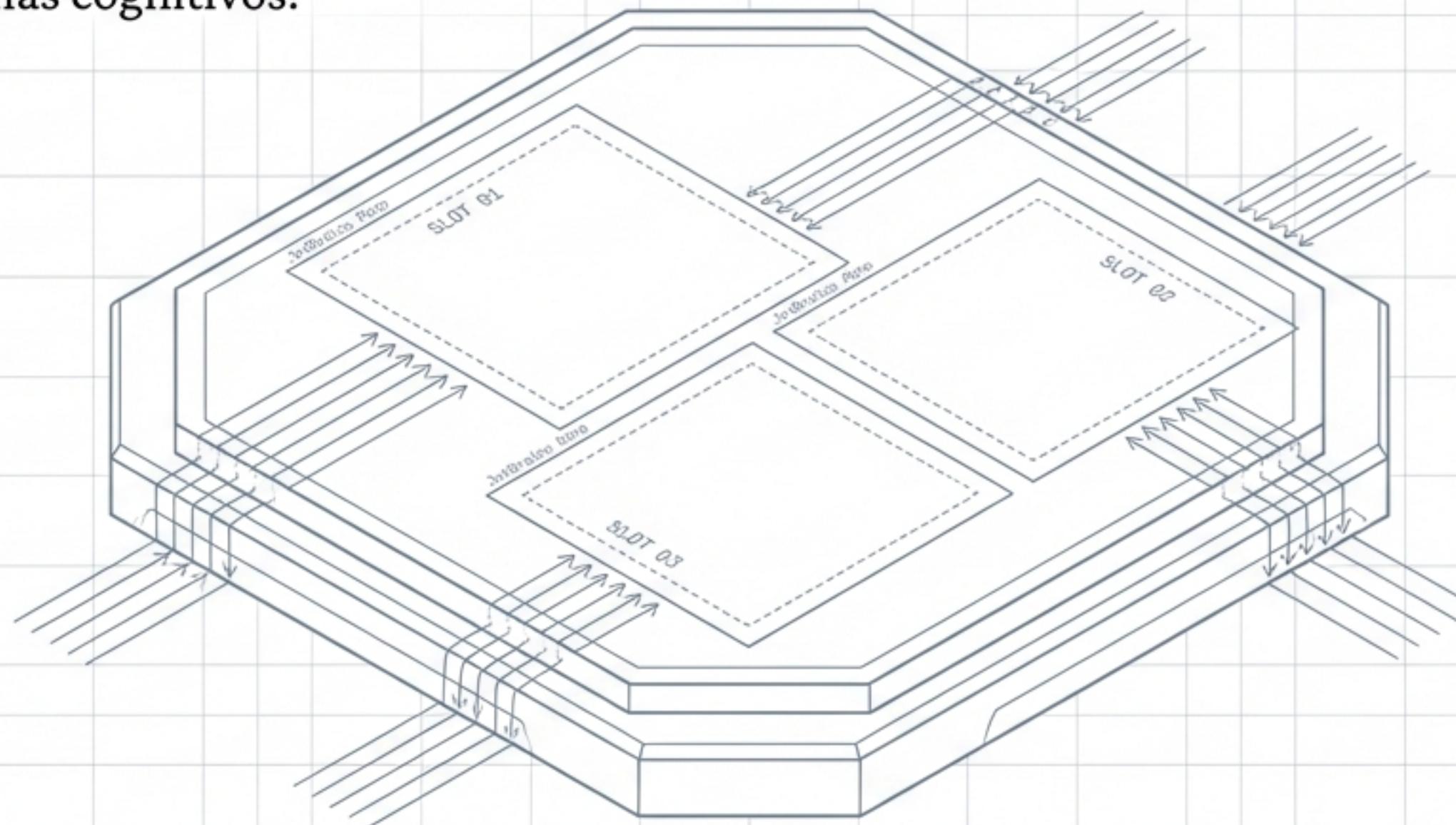


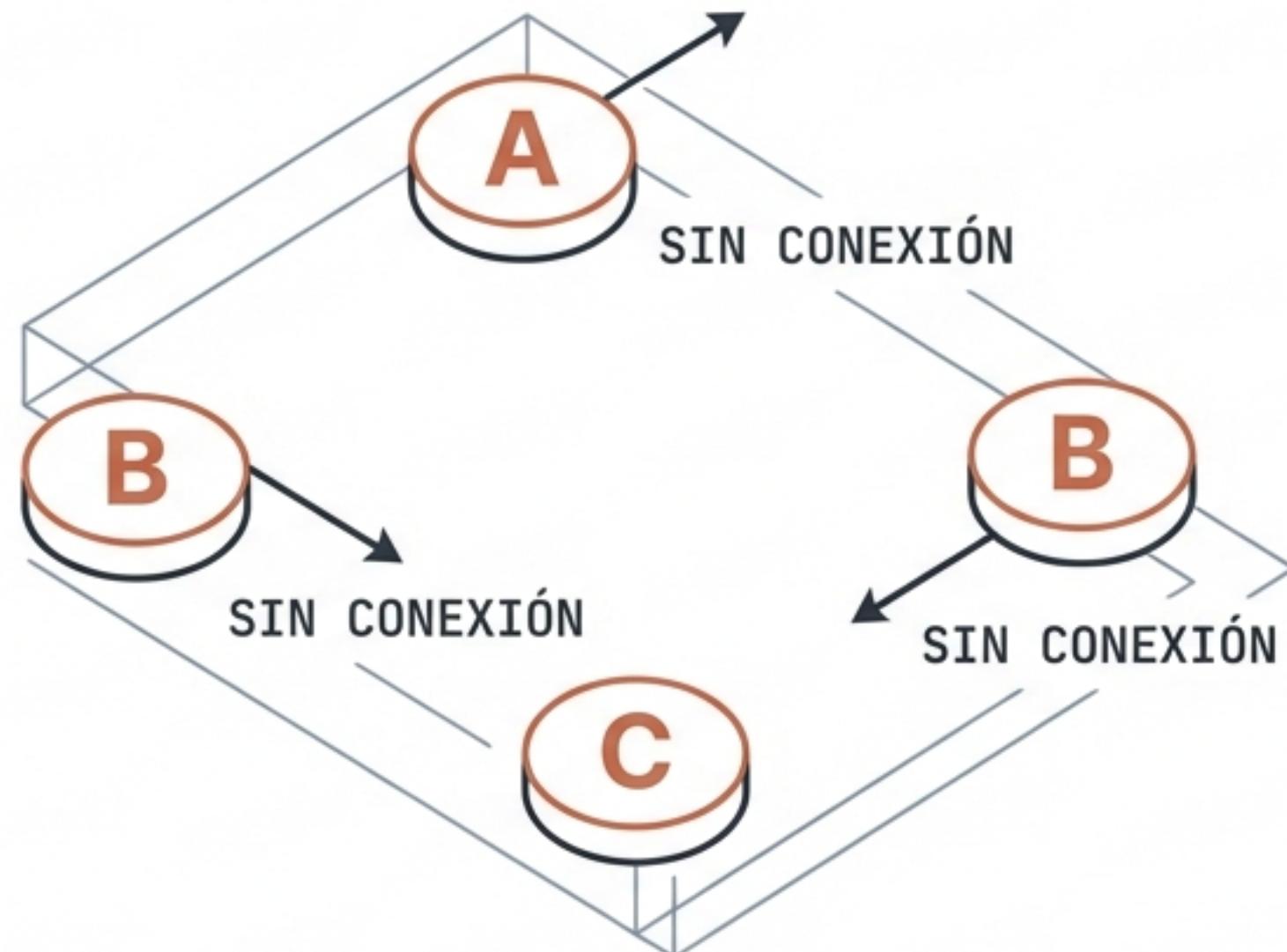
Arquitectura de Memoria para Agentes de IA

Implementando Memoria de Trabajo, Episódica y Semántica en sistemas cognitivos.

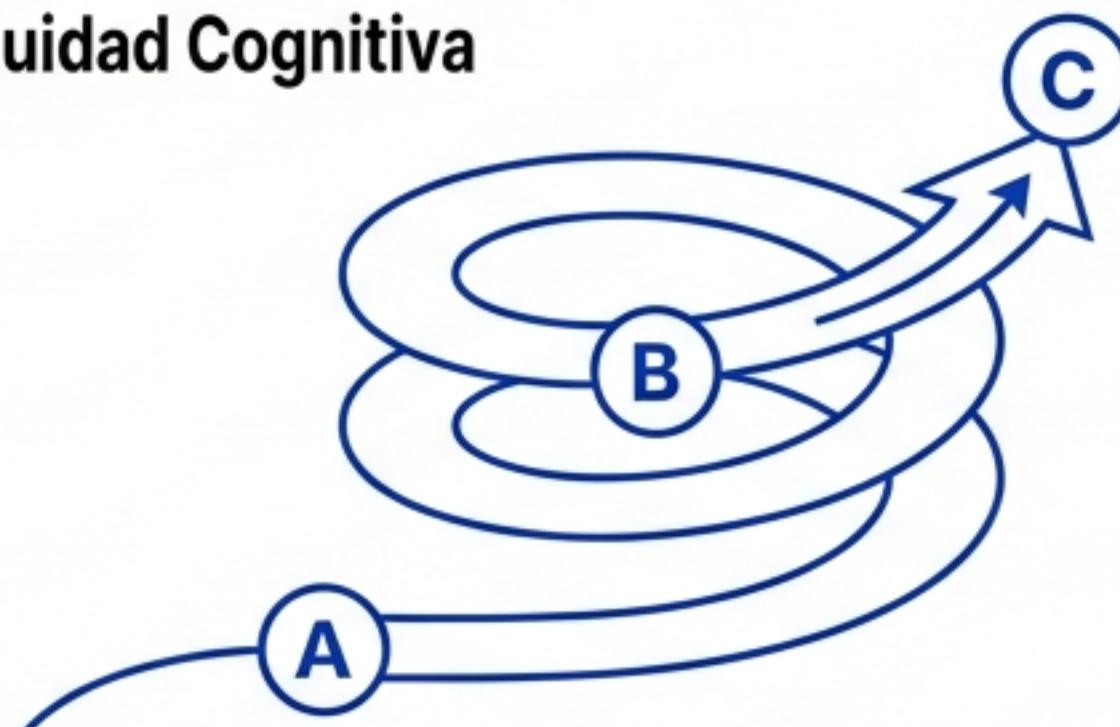


El desafío del agente sin estado

Bucle Desconectado



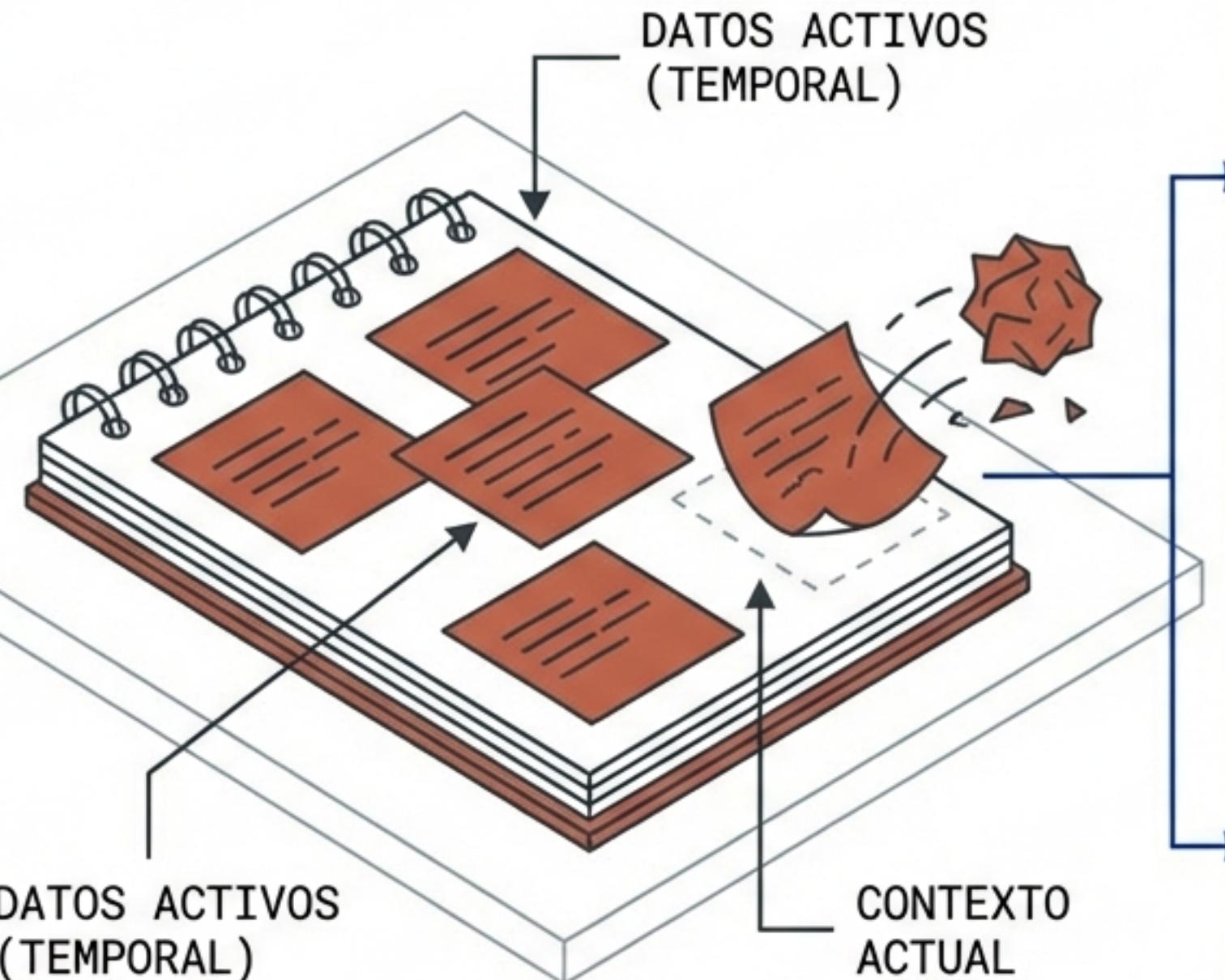
Continuidad Cognitiva



Sin una arquitectura de memoria, el agente vive en un eterno presente.

- **Pérdida de Hilo:** Incapacidad para mantener una conversación coherente.
- **Aislamiento de Acciones:** Los resultados de acciones previas no informan las decisiones futuras.
- **Razonamiento Limitado:** Falta de información temporal necesaria para deducir estados.

Memoria de Trabajo: El "Bloc de Notas" Mental



DEFINICIÓN

Componente diseñado para mantener el contexto **durante** una sesión de interacción activa. Al igual que un humano usa notas adhesivas, el agente necesita un espacio temporal para los datos activos.

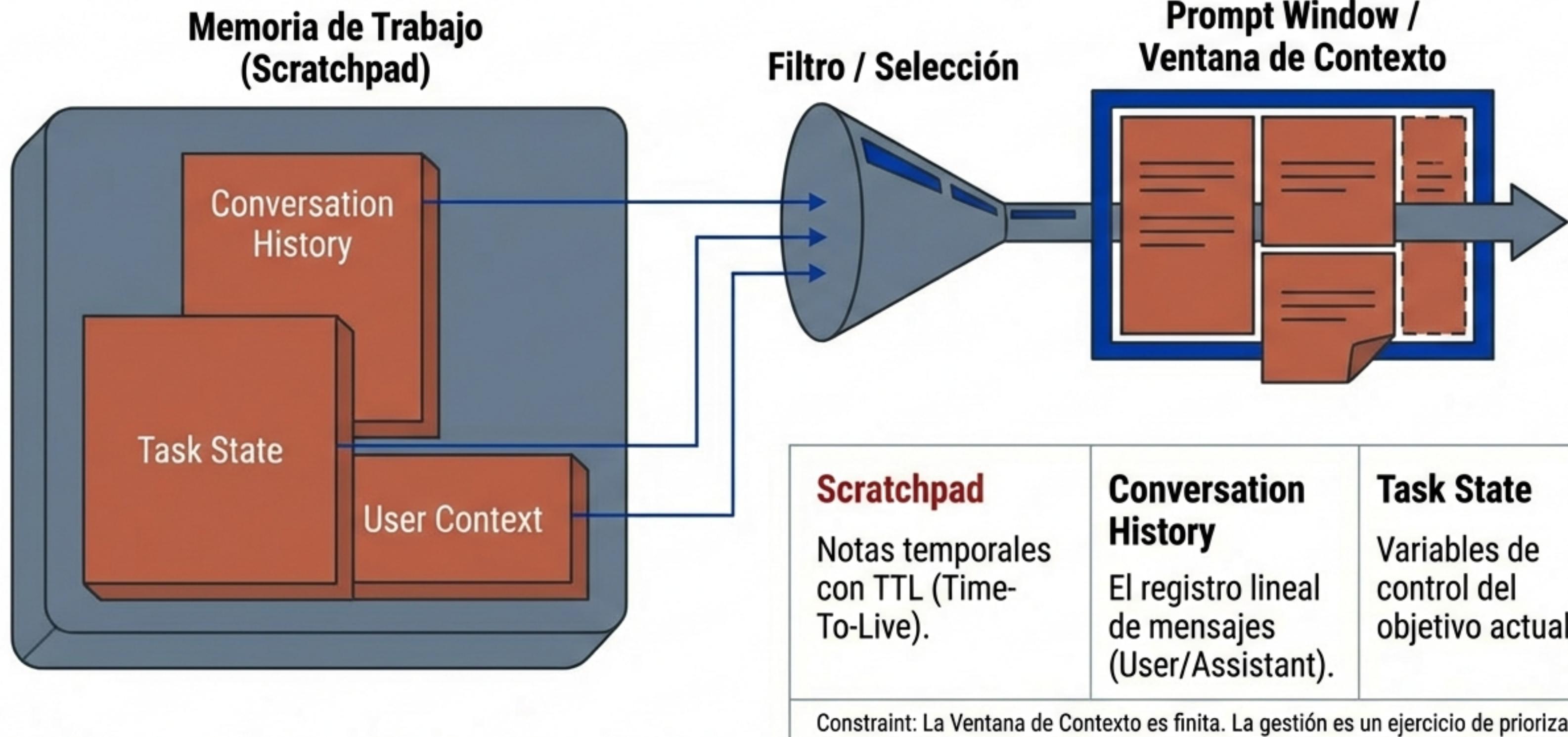
FUNCIONES CLAVE

Contexto de Usuario:
Preferencias e información inmediata.

Estado de la Tarea:
¿Cuál es el objetivo actual y cuánto hemos progresado?

Resultados Inmediatos:
Salidas de acciones recientes necesarias para el siguiente paso.

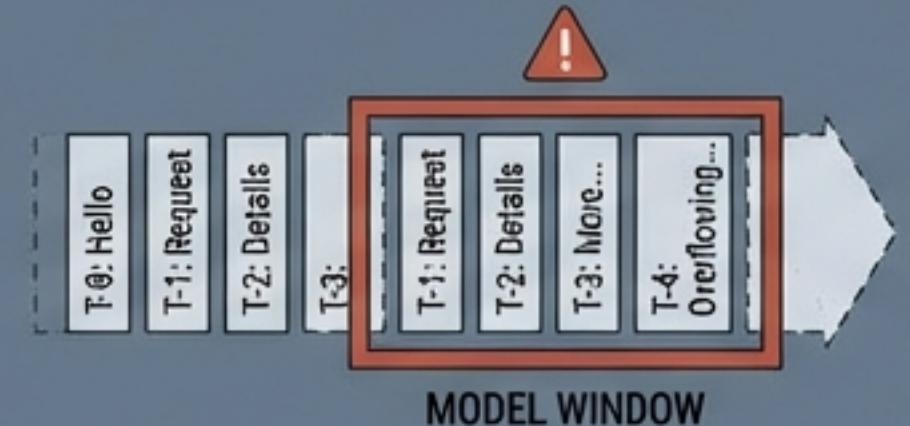
Arquitectura del Contexto Inmediato



Gestión de Contexto: Errores Comunes y Soluciones

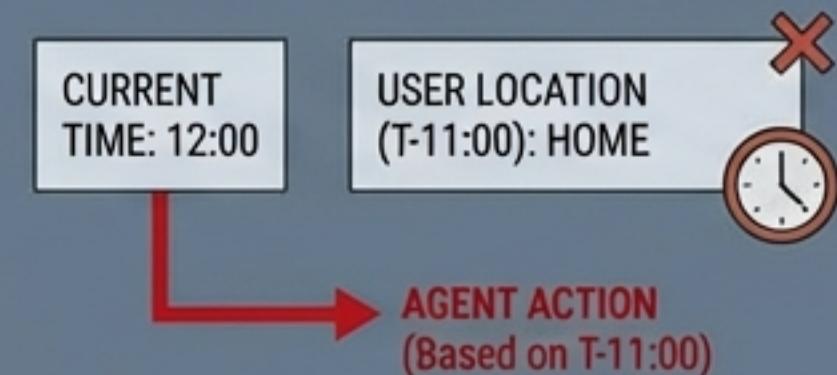
Error: Contexto Desbordado

El historial es más largo que la ventana del modelo.



Error: Memoria 'Stale' (Caducada)

El agente actúa con datos que ya no son ciertos.



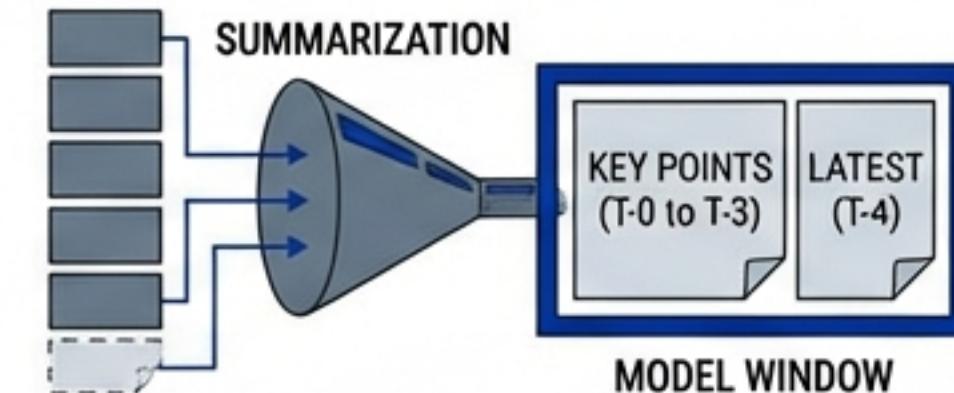
Error: Pérdida de Foco

Información crítica se borra al limpiar el buffer.



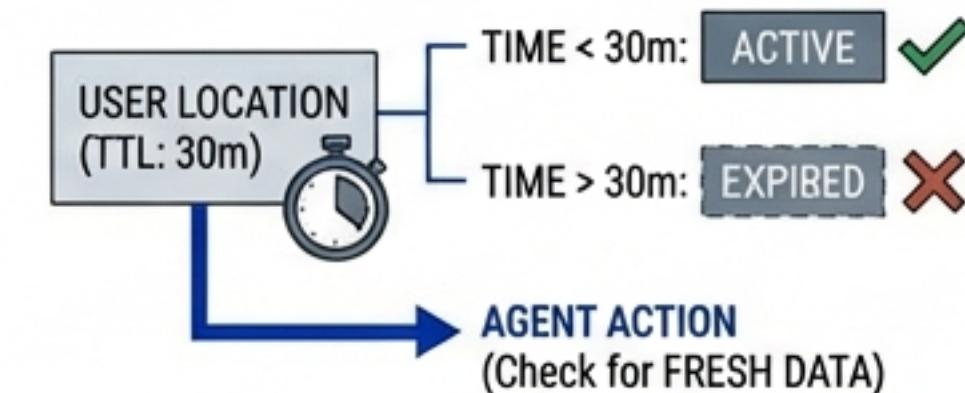
Solución: Resumir

Comprimir mensajes antiguos manteniendo los puntos clave.



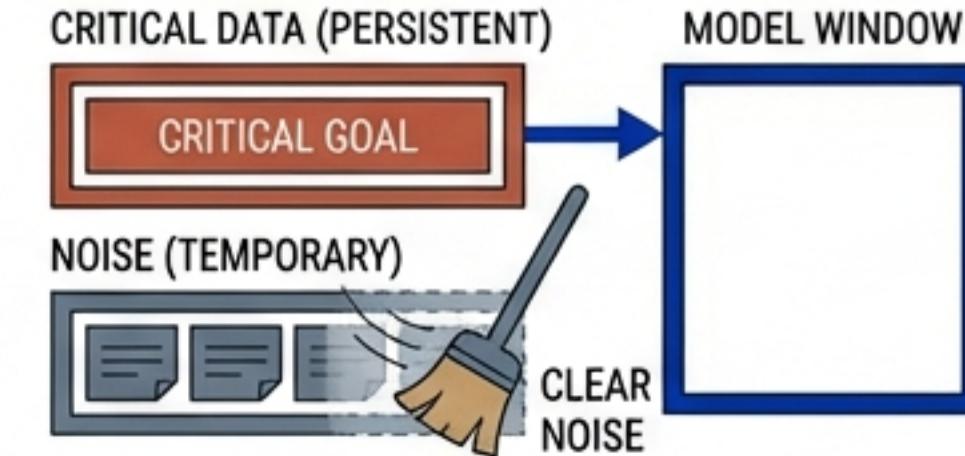
Solución: TTL (Time-To-Live)

Asignar caducidad a datos temporales.

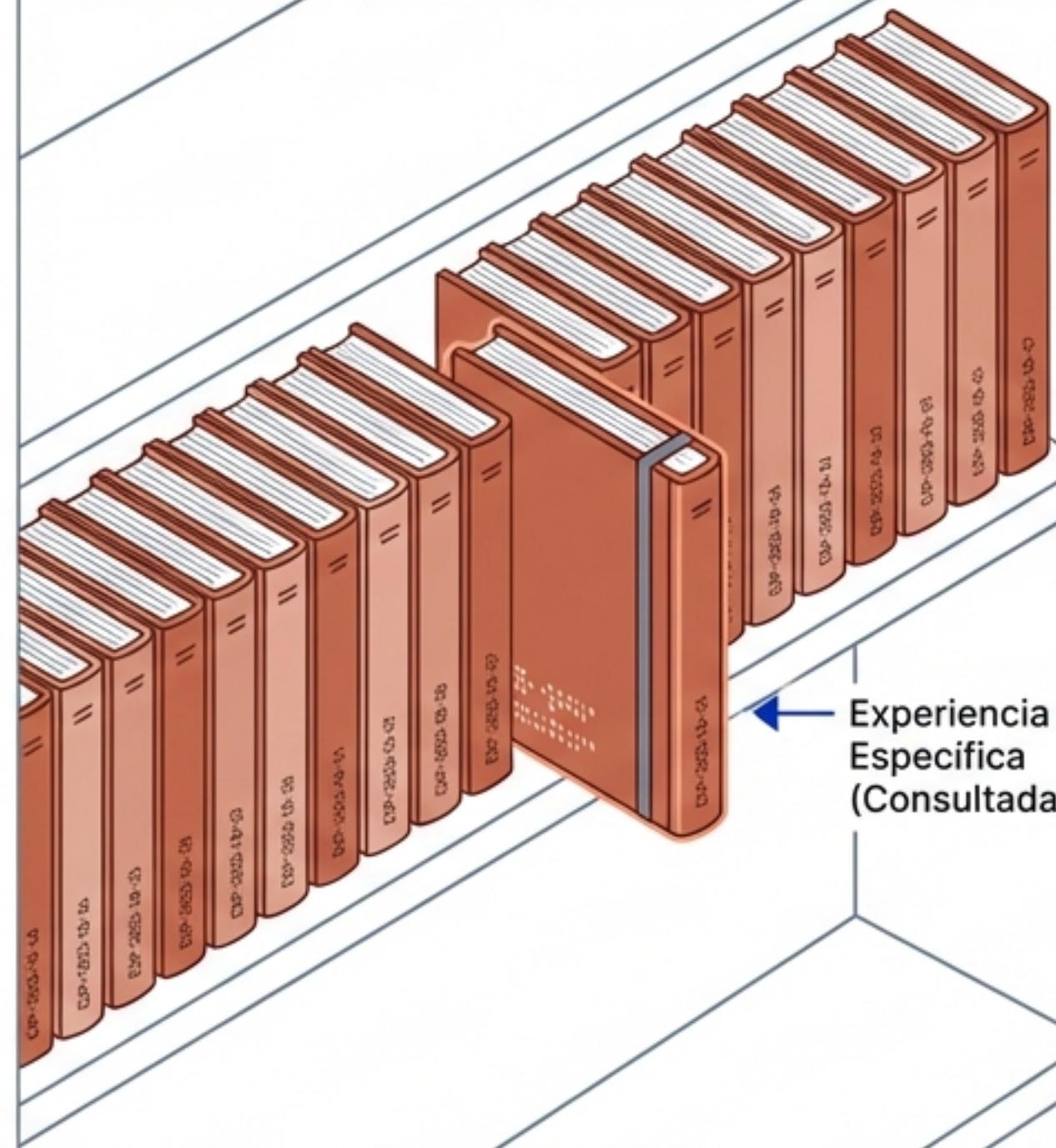
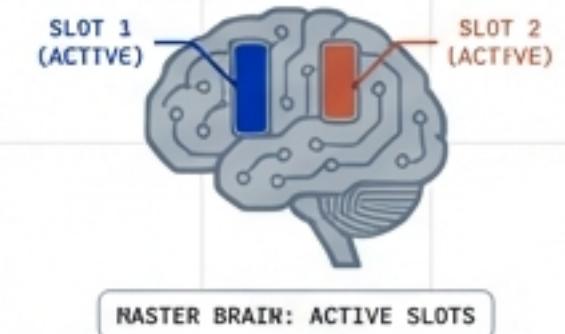


Solución: Priorización

Separar "Datos Críticos" (persistentes) de "Ruido".



Memoria Episódica: El "Diario" de Experiencias



Analogy: The Bookshelf of Past Experiences

Capacidad de recordar experiencias pasadas más allá de la sesión actual.

Un registro autobiográfico de experiencias específicas consultable para informar el futuro.

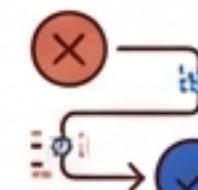
Types of Episodes



INTERACTION:
Diálogos previos con el usuario.



TASK:
Tareas completadas anteriormente.



ERROR:
Fallos pasados y resoluciones (lecciones).



INSIGHT:
Observaciones sobre patrones.

El Ciclo de la Experiencia: Store & Recall



Nota Técnica: Uso de Embeddings para encontrar contextos relacionados conceptualmente, no solo por palabras clave.

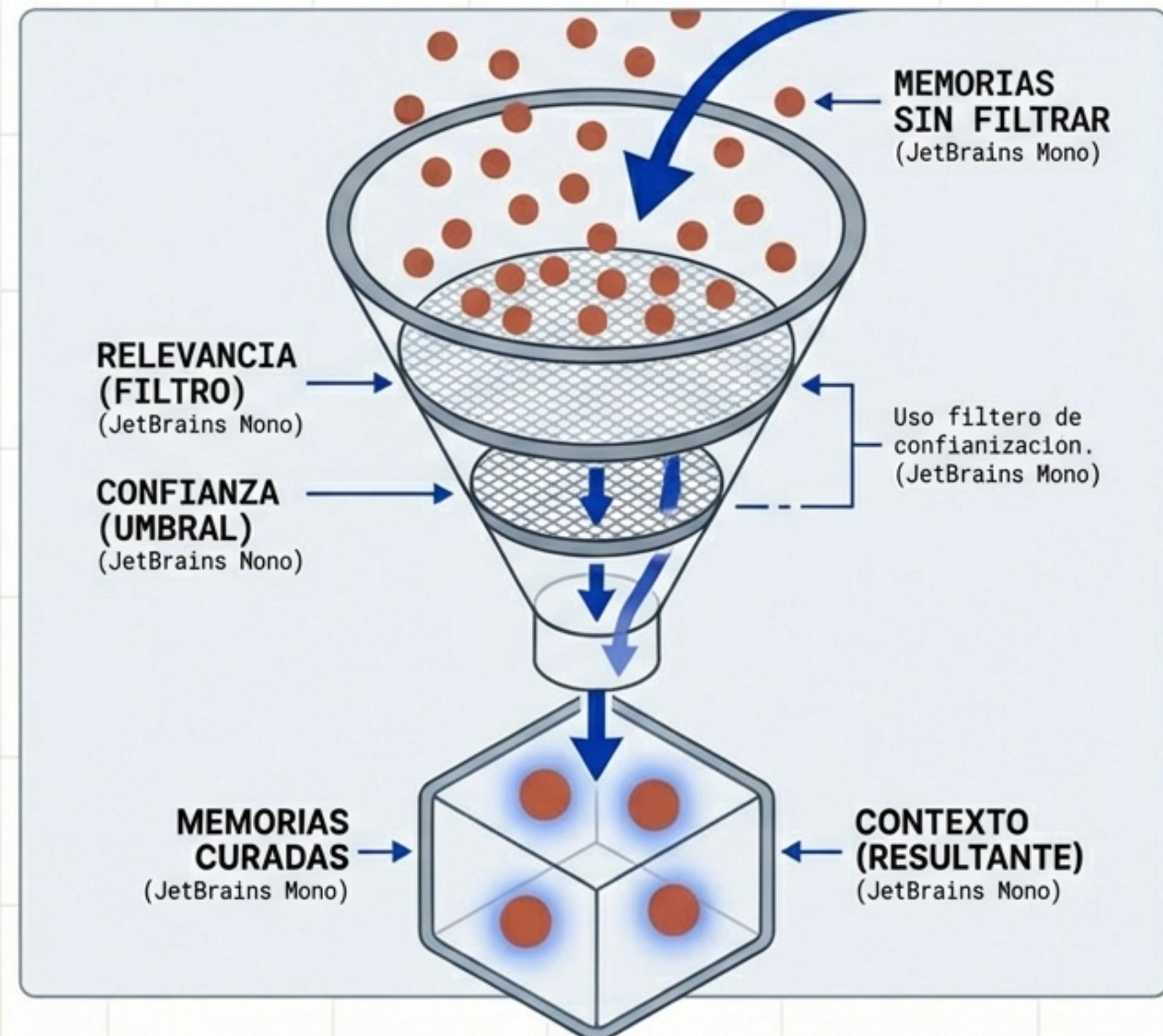
Curación de Recuerdos y "Olvido" Estratégico

La memoria infinita sin gestión conduce a ruido y alucinaciones.

Estrategias:

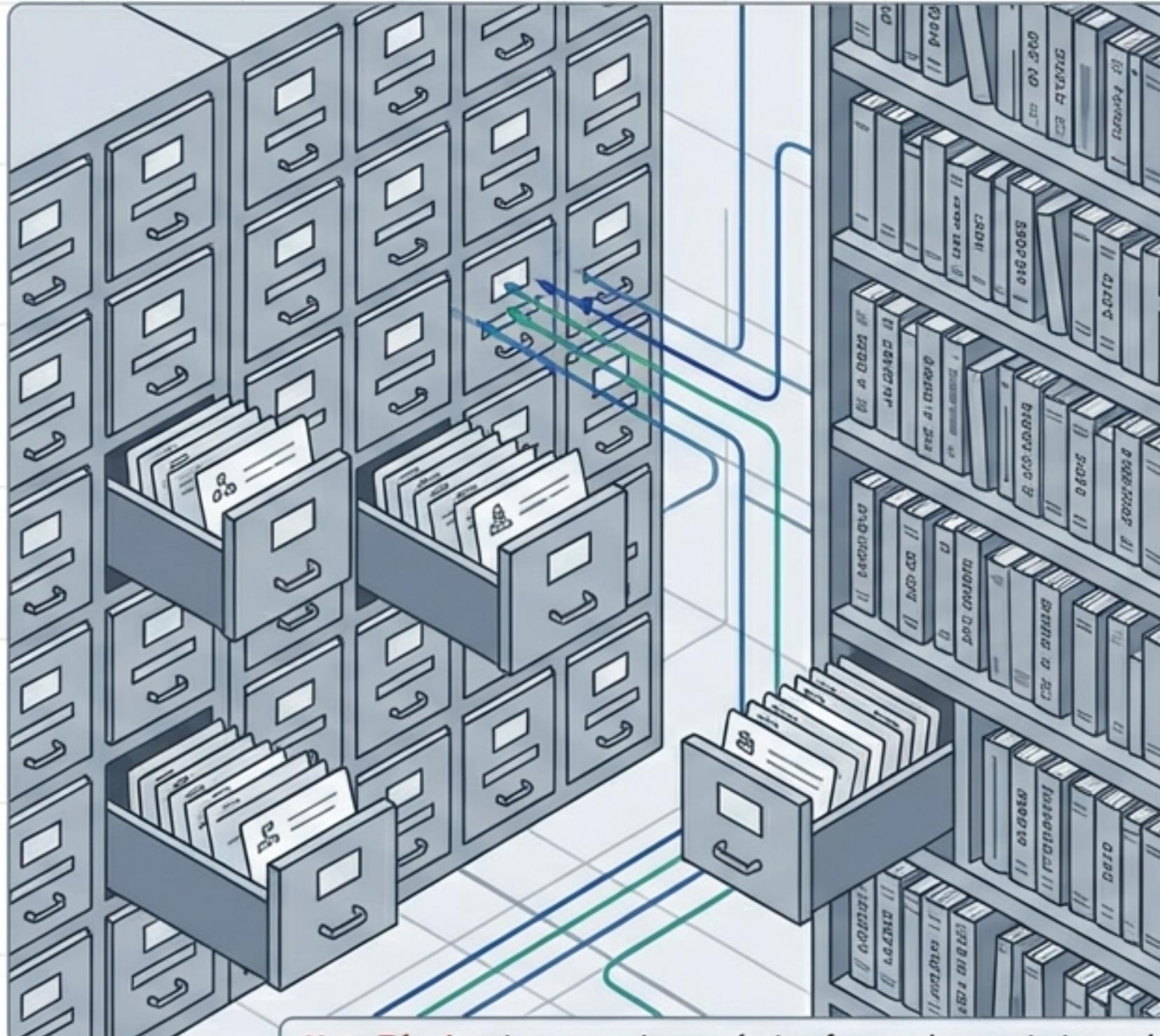
- **Filtrado por Relevancia:** Establecer un umbral mínimo de similitud para recuperar un recuerdo. 
- **Recency vs. Importance:** Ponderar recuerdos recientes frente a recuerdos fundacionales. 
- **Gestión de Alucinaciones:** Indicar 'Confianza' en recuerdos antiguos para evitar presentar datos obsoletos. 
- **Consolidación:** Fusionar múltiples episodios similares en una sola lección generalizada. 

Analogía Técnica: El embudo de la curación de la memoria.





Memoria Semántica: La "Enciclopedia" Interna



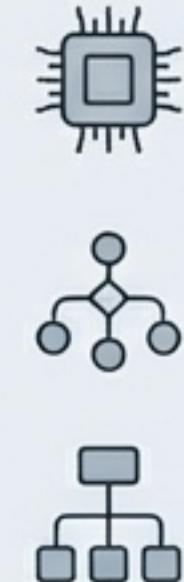
Nota Técnica: La memoria semántica forma el conocimiento del mundo, utilizado como base para la interpretación de nuevos datos.

Definición:

Almacenamiento de conocimiento factual generalizado, desvinculado de la experiencia específica. Es una biblioteca de referencia sobre cómo funciona el mundo.

Ejemplos:

- **Hechos:** "Python es un lenguaje de programación." (JetBrains Mono)
- **Relaciones:** "Machine Learning es un subconjunto de la IA."
- **Definiciones:** Ontologías y taxonomías del dominio.



Estructurando el Conocimiento: El Grafo



Nota Técnica

Estructura Tripla: Subject -> Predicate -> Object

**Conceptos
(Nodos):**

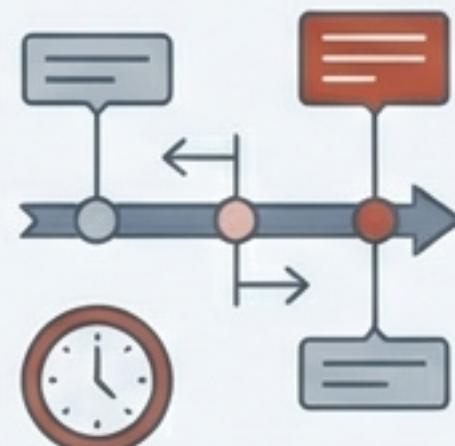
Entidades con definiciones.

**Relaciones
(Aristas):**

Conexión lógica.

Diferenciando Tipos de Memoria a Largo Plazo

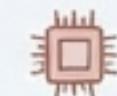
Memoria Episódica (El Pasado)



Naturaleza:
Autobiográfica.



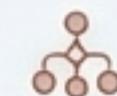
Contenido: “Ayer
compilé este código
y falló.”



(JetBrains Mono)



Fuente: Experiencia
directa del agente.

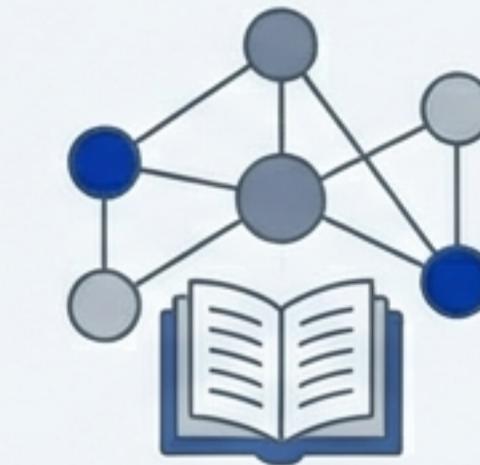


Memoria Semántica (El Conocimiento)

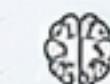
Naturaleza:
Factual / Abstracta.

Contenido: “Los errores
de compilación se deben
a sintaxis incorrecta.”

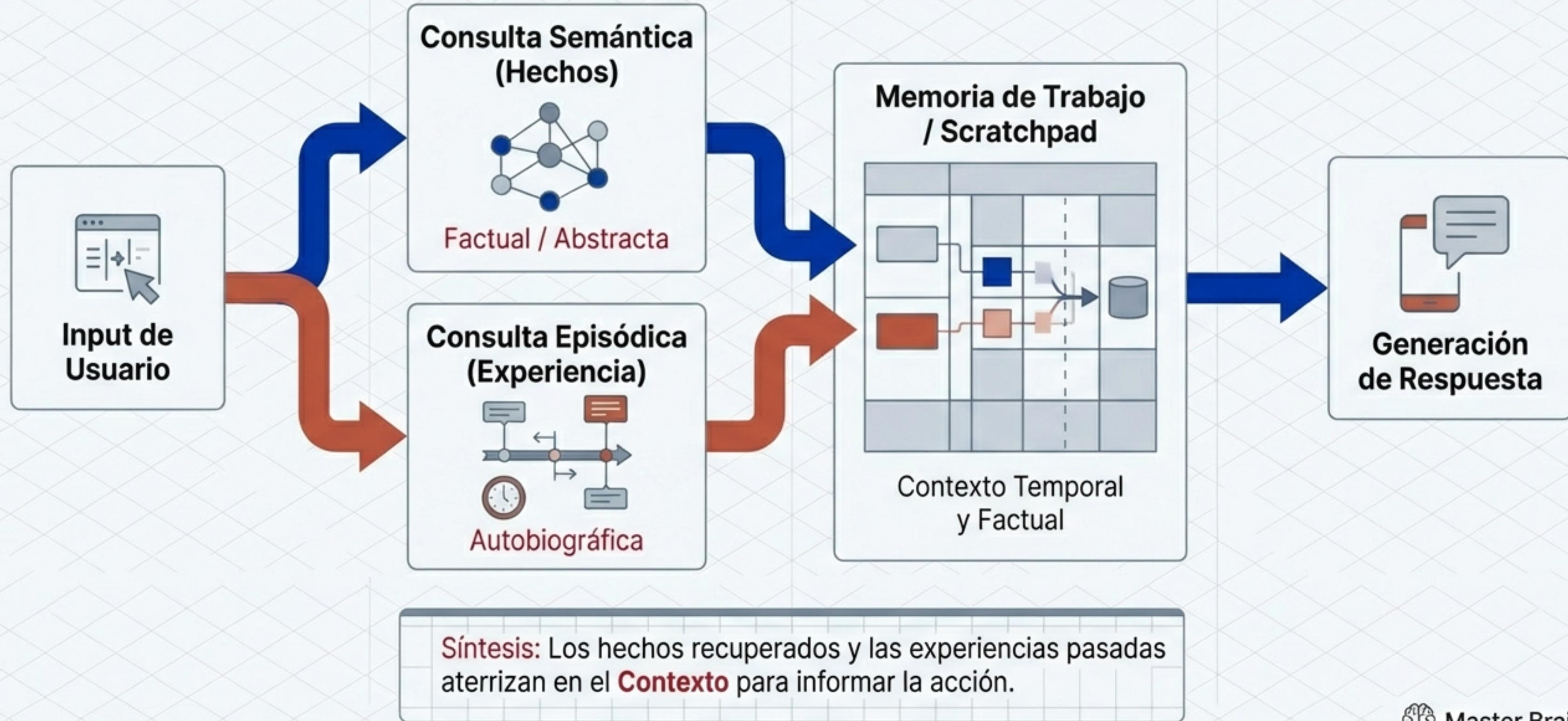
(JetBrains Mono)



Fuente: Bases de
conocimiento,
documentación.



El Cerebro Completo: Flujo de Integración



Aplicaciones Prácticas por Rol de Agente

Chatbot



Alta Memoria de Trabajo:
Historial de chat.



Episódica: Preferencias
de usuario.

Agente de Código

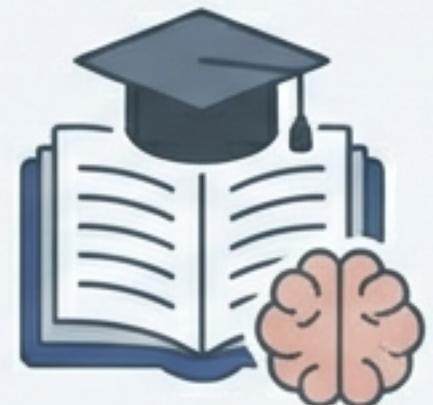


Alta Memoria Semántica:
Docs de lenguajes.

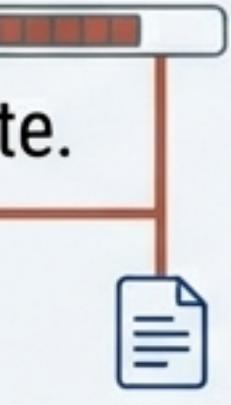


Trabajo:
Variables en scope.

Tutor AI



Alta Episódica:
Progreso del estudiante.

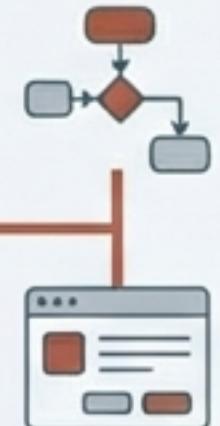


Semántica:
Materia a enseñar.

Soporte Técnico



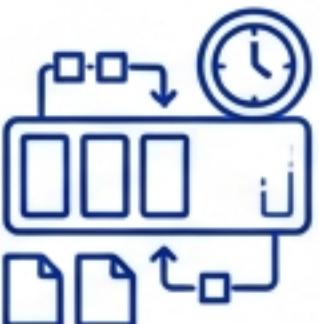
Episódica: Intentos
previos de solución.



Trabajo:
Contexto
del ticket actual.

Resumen Ejecutivo de Arquitectura

TRABAJO (Short-term)



Rol: Contexto Inmediato.

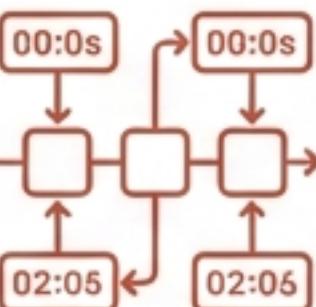


Key Tech: Prompt Engineering, Context Window.



Best Practice: TTL y Priorización.

EPISÓDICA (Long-term)



Rol: Experiencia Pasada.



Key Tech: Vector Stores, Embeddings.



Best Practice: Filtrado de relevancia.

SEMÁNTICA (World)



Rol: Hechos y Conocimiento.

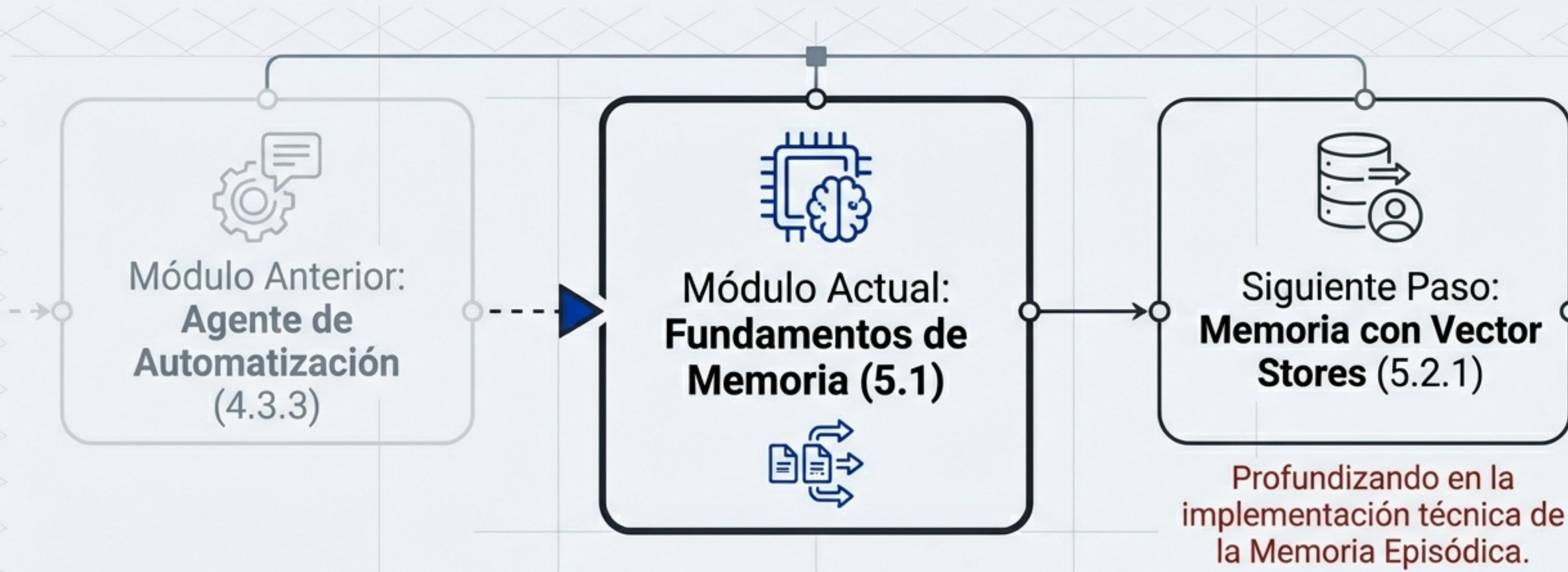


Key Tech: Knowledge Graphs, Ontologías.



Best Practice: Triplas Subject-Predicate-Object.

Ruta de Aprendizaje



“La memoria transforma un script de ejecución en un sistema capaz de aprender.”