

Multi-agente en producción

Coste, latencia, fallbacks y monitor.

CONCEPTOS CLAVE

- Arquitectura productiva: orquestador + subagentes especializados + capa MCP + monitor.
- Cinco **fallbacks** no negociables: timeout, low-score RAG, low-confidence LLM, loop detector, budget por sesión.
- Tres errores frecuentes: loops entre subagentes, sin presupuesto por sesión, tools sin JSON Schema.
- Observabilidad: *faithfulness*, p95 latencia, coste medio sesión, fallbacks/día.

EJEMPLO A MANO

Cuánto cuesta una petición compleja (asistente comercial BBVA):

$$\underbrace{0,0001}_{\text{routing}} + \underbrace{0,001}_{\text{RAG}} + \underbrace{0,0005}_{\text{tool}} + \underbrace{0,012}_{\text{LLM}_1} + \underbrace{0,008}_{\text{LLM}_2} + \underbrace{0,0001}_{\text{respuesta}} = \mathbf{0,022 \$}.$$

Volumen: 5 000 peticiones/día × 22 días laborables = 110 000 peticiones/mes. Coste mensual: ~ 2 420 \$ en LLMs. Latencia end-to-end: 40 + 280 + 350 + 520 + 460 + 80 = **1 730 ms**.

CONCEPTO → APLICACIÓN: CINCO PATRONES DE FALLBACK

Patrón	Cuándo activarlo	Acción
Timeout subagente	Subagente > 30 s.	Cancelar y devolver parcial.
Score bajo RAG	Top-1 chunk score < 0,4.	"No encontré información".
Confianza baja LLM	Logprob promedio < -3.	Escalar a humano (HITL).
Loop detectado	Misma tool 3 veces seguidas.	Detener y avisar al monitor.
Coste explotado	Sesión > 0,50 \$.	Resumir y devolver lo mejor.

Idea clave del Día 17. La **observabilidad determina si el asistente vive o muere**. Sin dashboard de *faithfulness*, p95, coste y fallbacks/día, todo proyecto agéntico se degrada en silencio.

Para llevarte: lectura Anthropic, *Building robust agentic systems*; práctica: orquestador con los cinco fallbacks implementados.