

Matriz de decisión IA

Modelo × Despliegue (M×D)

Una plantilla de una página para llevar a tu próximo comité. Antes de aprobar cualquier iniciativa de IA, sitúala en la matriz y respóndela con la checklist.

Los dos ejes

Eje M — **Modelo**: qué inteligencia usas, según criticidad de la tarea, complejidad y transparencia.

Eje D — **Despliegue**: dónde se ejecuta el agente, según sensibilidad del dato, costo y control.

La matriz 2×2

	Modelo tier medio (barato, predecible)	Modelo de frontera (potente, caro)
Despliegue local / on-device	Tareas recurrentes con dato sensible o regulado. Bajo costo marginal. <i>Ej.: clasificación diaria de documentos financieros.</i>	Trabajo complejo con dato sensible que no sale del perímetro. Vigila costo y hardware. <i>Ej.: análisis interno de código propietario.</i>
Despliegue en la nube	Alto volumen, dato público o interno, presupuesto ajustado. <i>Ej.: atención de primer nivel, borradores.</i>	Frontera intensiva en ráfagas, dato no crítico. Presupuestar por tokens. <i>Ej.: migración puntual de código.</i>

Regla de oro: paga potencia solo donde rinde y pon el dato sensible donde el control lo permita. La mayoría del trabajo operativo vive en la celda destacada.

Checklist de 6 preguntas

- 1 Tarea y riesgo** — ¿Cuál es la tarea concreta y qué cuesta un error (reputación, dinero, cumplimiento)?
- 2 Sensibilidad del dato** — ¿Es público, interno o regulado?
- 3 Despliegue** — ¿Local, nube o híbrido? ¿Por qué?
- 4 Tier de modelo** — ¿Qué tier pide la criticidad, no el hype?
- 5 Transparencia** — ¿El caso cae en un área con salvaguarda? ¿Cuál es el desempeño real ahí?
- 6 Gobernanza** — ¿Qué controles (permisos, auditoría, presupuesto) la hacen gobernable?

Cómo leer un benchmark sin comprar el titular

- ¿Quién corrió la prueba: el fabricante o un evaluador neutral?
- ¿Con qué andamiaje? Un mismo modelo rinde distinto según el harness.
- ¿El escenario de la prueba se parece a tu trabajo real?
- ¿Hay salvaguardas que bajen el desempeño en tu área de uso?