

L U M B R E



El precio de la IA: tokens, latencia y a qué le dices que no

Materia: Desarrollo de Software con IA

07/10/2026

Antes de empezar

Cada llamada a la IA cuesta: centavos por aquí, facturas enormes por allá si algo se descontrola. La moneda es el token —fragmentos de texto que entran y salen— y detrás vienen la latencia (segundos que el usuario espera) y la superficie de riesgo (tu clave, tus datos, la respuesta que se va de la app). Este OA te da la hoja de cálculo mental para estimar el costo de una función antes de construirla, los cuatro candados técnicos innegociables y cuándo responder «aquí no va IA».

QUÉ APRENDERÁS

- Estimar costo por uso, controlar secretos y decidir cuándo NO poner IA.

Cada «hola» cuesta: la economía de los tokens

La IA no cobra por pregunta: cobra por **tokens** —pedazos de texto, aproximadamente tres cuartos de palabra— que entran (tu prompt, los documentos pegados) y salen (la respuesta). Una petición RAG con cinco fragmentos pegados puede costar cien veces más en entrada que un chat simple; una función que el usuario pulsa sin saber que hay una llamada detrás puede multiplicar tu factura sin multiplicar tu valor. El builder que no sabe estimar no dirige: reza.

ACTIVIDAD INTERACTIVA DISPONIBLE EN
LUMBRE

1. La hoja de cálculo mental



Los cuatro candados técnicos (innegociables)

El orden es el de siempre en este curso: la técnica la monta la IA en dos prompts; lo que es tuyo es EXIGIRLOS. Ningún candado de esta tabla es opcional «por ahora».

Candado	Qué hace	Sin él...
La clave vive del lado del servidor	la app llama a TU backend; el backend llama a la IA con la clave guardada en variable de entorno	cualquiera ve tu clave en el navegador y la gasta (o quema) a tu cargo
Límite de tokens de salida	tope duro en la petición: la respuesta no puede pasar de N palabras	un modelo enganchado en un bucle puede escribir hasta que tu tarjeta diga basta

Candado	Qué hace	Sin él...
Caché de respuestas repetidas	la misma pregunta frecuente se responde del caché, no de la IA	pagas por lo que ya contestaste diez veces hoy
Presupuesto con alarma	techo mensual de gasto con aviso temprano al 50 y 80 %	la factura del mes que viene es tu primera noticia del descontrol

El prompt del presupuesto antes del código

Números, no adjetivos: la decisión de seguir es después de la cuenta

Antes de construir la función de resumen de tickets con IA, hazme la hoja de costos:\n\n1. ESTIMACION: tokens aproximados de entrada y salida por llamada con mi diseño actual\n (pego el hilo completo). Precio por llamada y por 1.000 usuarios al mes.\n2. OPTIMIZACION: proponme 3 variantes mas baratas (recortar hilo, resumen parcial,\n cachear por plantilla) con su costo y su pérdida de calidad.\n3. CANDADOS: confirma en el diseño los 4 innegociables: clave solo en servidor con\n variable de entorno, límite de tokens de salida, caché de repetidas, y techo\n mensual con alarma al 50 y 80 %.\n4. LATENCIA: tiempo esperado de respuesta; si pasa de 3 segundos, propón como\n mostrar avance al usuario (streaming o indicador).\n\nDame números, no adjetivos. Si la cuenta sale cara para el valor que entrega,\ndilo en una línea y yo decido si sigue.

Para reforzar

Prueba

2. Tu app tiene un botón «resumir este ticket» que llama a la IA. ¿Qué diseño es el correcto?

- ☐ La clave de la IA en el JavaScript del navegador, para que la respuesta llegue rápido.
- ☐ El navegador llama a tu backend; tu backend llama a la IA con la clave en variable de entorno y límite de tokens de salida.
- ☐ La clave en el frontend pero escondida en un archivo de configuración ofuscado.
- ☐ Cada usuario registra su propia clave para no gastar la tuya.

3. Una función con IA es cara solo si la respuesta es larga; la entrada del prompt casi no pesa en el costo.

☐ Verdadero

☐ Falso

Detecta el «IA» que no debía ser IA

ACTIVIDAD
INTERACTIVA
DISPONIBLE EN
LUMBRE

4. Cuatro decisiones de funciones en proyectos de alumnos. Señala la que NO merece IA:



5. Toma la función con IA que estés considerando (o la del OA anterior) y hazle la hoja de costos del prompt: números por llamada, por mes, con las tres variantes de ahorro y los cuatro candados marcados uno a uno. Termina con la decisión honesta en una línea: sigue, se recorta, o aquí no va IA.

Saber qué construir incluye saber qué NO construir: la cuenta es la que manda.

Referencia curricular

Materia:	Desarrollo de Software con IA
-----------------	-------------------------------

Versión interactiva completa (con actividades, videos y autoevaluación):



<https://empresa.lumbre.lat/object/fsi-m20-0a03>

Material educativo generado desde Lumbre · lumbre.lat