La unidad de medida del cerebro: qué es un token

Datos de esta lección

Fase	F1 — Cimientos, seguridad y arquitectura del sistema
Módulo	M1.2 — Modelos mentales: radiografía de un agente telefónico
Puntos de profesionalidad al superar el test	15

Objetivo práctico

Al terminar esta lección podrás explicar qué es un token, por qué los LLM cobran y miden su uso en tokens, y por qué esto importa directamente para calcular el coste de una llamada telefónica con IA.

Resultado que obtendrás

Comprensión suficiente de qué es un token para seguir, sin sorpresas, el módulo 2.1 («Gestión financiera y eliminación de riesgos»), donde se calcula el coste real de una llamada.

Dónde encaja dentro del sistema final

Cada palabra que el LLM lee (del prompt, de la conversación) y cada palabra que genera como respuesta consume tokens, y los proveedores de LLM cobran según la cantidad de tokens usados. Es la unidad económica básica del «cerebro» del sistema.

Conocimientos previos

- LLM y prompt (lecciones anteriores de este módulo).

Herramientas necesarias

- Un navegador con NVDA activo, en modo de exploración.

Posible coste

Esta lección es conceptual. El cálculo real de coste por token se practica con cifras concretas en el módulo 2.1.

Modelo mental

Piensa en los tokens como en los minutos de una llamada telefónica tradicional: la compañía no te cobra por «una conversación», te cobra por la duración medida en una unidad concreta (minutos). Un LLM no cobra por «una respuesta», cobra por la cantidad de texto procesado y generado, medida en una unidad concreta: el token.

Explicación

Un token es la unidad mínima de texto que un LLM procesa internamente. No equivale exactamente a una palabra: puede ser una palabra completa, parte de una palabra larga, o un signo de puntuación, dependiendo del idioma y del modelo. Como referencia aproximada y no exacta, en español un token equivale, de media, a algo menos que una palabra completa.

Los proveedores de LLM (OpenAI, Anthropic, Google, entre otros) cobran normalmente según dos cantidades: los tokens de *entrada* (lo que se le envía al modelo: el prompt más la conversación hasta ese momento) y los tokens de *salida* (lo que el modelo genera como respuesta). Cuanto más largo es el prompt y más larga es la conversación acumulada, más tokens de entrada se procesan en cada turno, lo cual tiene un impacto económico directo en un agente telefónico que mantiene conversaciones largas. Este dato se retoma con cifras y calculadoras reales en el módulo 2.1.

Vocabulario nuevo

Token

Piénsalo así: El minuto de una llamada telefónica: la unidad por la que se mide y se cobra el uso.

Unidad mínima de texto que procesa un LLM; puede ser una palabra, parte de una palabra o un signo de puntuación.

Ejemplo: Una frase de diez palabras en español puede equivaler aproximadamente a entre doce y quince tokens, dependiendo del modelo.

Preparación

No se requiere preparación técnica.

Procedimiento paso a paso

1. **Contexto:** Vas a localizar «Token» en el Glosario del curso.

   Acción de teclado: Abre el Glosario (`/glosario.php`) y pulsa `H` hasta el encabezado de nivel 3 «Token», en el bloque «T».

   Respuesta esperada de NVDA: NVDA debería anunciar «Token, encabezado de nivel 3».

   Qué significa: Confirmas que el término está documentado con la analogía del minuto telefónico y su definición técnica.

   Acción siguiente: Lee la definición completa con la flecha abajo.

2. **Contexto:** Vas a comprobar, usando el Buscador del curso, dónde se retoma este concepto más adelante con cifras reales.

   Acción de teclado: Ve al Buscador (`/buscar.php`), escribe `token` en el campo «Término de búsqueda» y pulsa `Enter`.

   Respuesta esperada de NVDA: NVDA debería anunciar el número de resultados y, entre ellos, la lección «Cómo se calcula el coste real de una llamada».

   Qué significa: Ese resultado corresponde al módulo 2.1, donde este mismo concepto se retoma con cifras y una calculadora real.

   Acción siguiente: No hace falta abrir esa lección todavía: basta con confirmar que aparece en los resultados.

Errores frecuentes y diagnóstico

Asumir que un prompt muy largo no tiene ningún coste adicional respecto a uno breve.

Cómo localizarlo: Un prompt más largo consume más tokens de entrada en cada turno de la conversación, lo que incrementa el coste por llamada.

Solución: Diseñar el prompt con la información necesaria pero sin relleno innecesario; el módulo 2.1 enseña a calcular este impacto con cifras reales.

Comprobación final

Explica con tus palabras por qué una conversación telefónica larga con un agente de IA puede costar más que una corta, en términos de tokens. Si tu explicación menciona que tanto el prompt como el historial de la conversación se reenvían en cada turno, la comprobación es positiva.

Qué acabas de conseguir

Tienes la base conceptual necesaria para entender, con cifras reales, el módulo 2.1 sobre gestión financiera y eliminación de riesgos.

Ejercicio práctico

Cuenta aproximadamente cuántas palabras tiene esta lección y estima, usando la referencia de esta lección (un token es algo menos que una palabra), cuántos tokens supondría procesarla como si fuera un prompt.

Resumen de lo imprescindible

- Un token es la unidad mínima de texto que procesa un LLM, no siempre igual a una palabra completa.
- Los LLM se cobran normalmente por tokens de entrada y de salida, de forma separada.
- Un prompt o una conversación más larga consume más tokens y, por tanto, tiene mayor coste.

No necesitas aprender todavía

Todavía no necesitas calcular con precisión el coste de tokens de ningún proveedor concreto: se hace con cifras reales en el módulo 2.1.

Estoy preparado para continuar si puedo…

- Explicar qué es un token usando la analogía del minuto telefónico.
- Explicar por qué una conversación más larga tiende a costar más.
- Distinguir tokens de entrada y tokens de salida.

Test de comprobación

> *Para registrar el resultado y obtener Puntos de profesionalidad, realiza este test desde la versión web de la lección. Las opciones se muestran aquí sin señalar la respuesta correcta.*

1. Dos agentes usan el mismo LLM. El Agente A tiene un prompt de 200 palabras; el Agente B tiene un prompt de 2.000 palabras con la misma calidad de instrucciones. En igualdad del resto de condiciones, ¿qué se puede afirmar sobre su coste por llamada?

    a. Ambos cuestan exactamente lo mismo, porque el coste no depende de la longitud del prompt.
    b. El Agente B tenderá a costar más por llamada, porque su prompt más largo se reenvía como tokens de entrada en cada turno de la conversación.
    c. El Agente A costará más, porque los prompts cortos consumen más tokens por palabra.

2. ¿Por qué un token no equivale exactamente a una palabra?

    a. Porque un token puede ser una palabra completa, parte de una palabra larga, o un signo de puntuación, según el idioma y el modelo.
    b. Porque un token equivale siempre exactamente a diez palabras.
    c. Porque los tokens sólo existen en inglés, nunca en español.

3. Un LLM cobra por separado los tokens de entrada y los de salida. ¿Qué representan, respectivamente, cada uno de ellos en la conversación de un agente telefónico?

    a. Entrada: el prompt más la conversación acumulada que se le envía al modelo. Salida: el texto que el modelo genera como respuesta.
    b. Entrada: sólo la primera frase de la llamada. Salida: sólo la última frase de la llamada.
    c. Entrada y salida significan exactamente lo mismo; es sólo cuestión de nomenclatura del proveedor.

Fuentes

- Token counting — Anthropic (consultado el 2026-08-30)

Última actualización de este contenido: **2026-02-10**.