

Día 11 · Tema 4 — Ejercicios: LLMs, sampling y coste

Cómo se usa esta hoja. Hay dos tipos de ejercicio, y sólo dos.

El ejercicio del día. Uno por hoja. Es el obligatorio y es el que entra en el examen, con el cálculo incluido. Basta con hacer ese para seguir el curso con solvencia.

Los ejercicios de ampliación. El resto. Están para enseñar de qué es capaz el método, no para exigir: **no se piden calculados en el examen.** De ellos sólo puede preguntarse la idea, a nivel conceptual.

EL EJERCICIO DEL DÍA *obligatorio · entra en el examen*

1. Temperature scaling

Un modelo de lenguaje genera, para la posición $t + 1$, los logits $\ell = (2,0; 1,0; 0,5; -1,0)$ correspondientes a cuatro tokens $\{A, B, C, D\}$.

La distribución con temperatura τ es

$$p_{\tau}(x_i) = \frac{\exp(\ell_i/\tau)}{\sum_j \exp(\ell_j/\tau)}.$$

- (a) Calcula la distribución para $\tau = 1,0$ (softmax estándar). Muestra los valores de $\exp(\ell_i)$ y la suma Z .
- (b) Calcula la distribución para $\tau = 0,5$. ¿Qué sucede con la probabilidad del token más probable?
- (c) Calcula la distribución para $\tau = 2,0$. ¿Qué sucede ahora?
- (d) Describe con una frase el efecto de $\tau < 1$ y de $\tau > 1$ sobre la distribución.

Pista. Valores útiles: $e^2 \approx 7,389$, $e^1 \approx 2,718$, $e^{0,5} \approx 1,649$, $e^{-1} \approx 0,368$, $e^4 \approx 54,60$, $e^{-2} \approx 0,135$, $e^{0,25} \approx 1,284$, $e^{-0,5} \approx 0,607$.

AMPLIACIÓN *opcional · no se pide calculado en el examen*

2. Top- k sampling

Partiendo de la distribución obtenida en el Ejercicio 1 con $\tau = 1,0$:

- (a) Aplica top- k con $k = 2$. ¿Qué tokens sobreviven al truncamiento?
- (b) Calcula las probabilidades renormalizadas de los tokens supervivientes.
- (c) ¿Qué porcentaje de la masa de probabilidad original se ha descartado?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

3. Top- p (nucleus) sampling

Aplica ahora el muestreo top- p con $p = 0,9$ a la distribución obtenida con $\tau = 0,7$ y los mismos logits del Ejercicio 1.

- (a) Calcula primero la distribución con $\tau = 0,7$. Usa $\ell_i/0,7$: los logits escalados son aproximadamente $(2,857; 1,429; 0,714; -1,429)$.
Pista. $e^{2,857} \approx 17,41$, $e^{1,429} \approx 4,173$, $e^{0,714} \approx 2,043$, $e^{-1,429} \approx 0,240$.
- (b) Ordena los tokens de mayor a menor probabilidad y acumula hasta alcanzar o superar $0,9$. ¿Qué tokens forman el nucleus?
- (c) Calcula las probabilidades renormalizadas del nucleus.

AMPLIACIÓN *opcional · no se pide calculado en el examen*

4. Cálculo de coste de API

Un chatbot de atención al cliente recibe 500 consultas al día. Cada consulta tiene:

Componente	Tokens
System prompt	800
Historial de conversación	1 200
Mensaje del usuario	100
Total input	2 100
Respuesta media (output)	300

El modelo utilizado es Claude Sonnet con precios $p_{\text{in}} = 3 \text{ \$/M tokens}$ y $p_{\text{out}} = 15 \text{ \$/M tokens}$. La fórmula de coste es:

$$C = \frac{N_{\text{in}}}{10^6} p_{\text{in}} + \frac{N_{\text{out}}}{10^6} p_{\text{out}}.$$

- (a) Calcula el coste por consulta.
- (b) Calcula el coste diario (500 consultas).
- (c) Calcula el coste mensual (30 días).

AMPLIACIÓN *opcional · no se pide calculado en el examen*

5. Ahorro con prompt caching

El mismo chatbot del Ejercicio 4 implementa *prompt caching*: el system prompt de 800 tokens se almacena en caché y se factura a un precio reducido de $0,30 \times p_{\text{in}} = 0,90$ \$/M tokens, en lugar del precio completo de 3 \$/M.

- (a) Calcula el nuevo coste de entrada por consulta, separando la parte cacheada (800 tokens a 0,90 \$/M) y la parte no cacheada (1 300 tokens a 3 \$/M).
- (b) Calcula el coste total por consulta (entrada + salida).
- (c) Calcula el ahorro mensual respecto al escenario sin caché del Ejercicio 4.
- (d) ¿Qué porcentaje de reducción supone sobre el coste mensual original?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

6. Presupuesto de contexto (context budget)

Un sistema RAG (*Retrieval-Augmented Generation*) opera con un presupuesto de contexto de $B = 4\,096$ tokens, distribuido así:

Componente	Tokens
System prompt	500
Historial de conversación	1 200
Consulta del usuario	150
Reserva para la respuesta	600
Total fijo	2 450

El presupuesto restante se destina a fragmentos recuperados de la base de conocimiento, cada uno de 400 tokens.

- (a) ¿Cuántos tokens quedan disponibles para fragmentos?
- (b) ¿Cuántos fragmentos completos caben?
- (c) El equipo de producto necesita incluir 6 fragmentos. ¿Cuántos tokens adicionales se requieren? Propón al menos dos estrategias concretas para liberar ese espacio.

AMPLIACIÓN *opcional · no se pide calculado en el examen*

7. Comparación de modelos

Usando las mismas cifras del chatbot (500 consultas/día, 2 100 tokens de entrada, 300 tokens de salida, 30 días/mes), calcula el coste mensual para cada uno de los siguientes modelos:

Modelo	p_{in} (\$/M)	p_{out} (\$/M)
Claude Sonnet	3,00	15,00
GPT-5	2,50	10,00
Mistral Large	2,00	6,00
Gemini 1.5 Pro	1,25	5,00

- (a) Construye una tabla con el coste mensual de entrada, de salida y total para cada modelo.
- (b) ¿Cuánto se ahorra al mes eligiendo el modelo más barato frente al más caro?
- (c) Comenta brevemente qué factores, además del precio, debería considerar la empresa al elegir modelo.

AMPLIACIÓN *opcional · no se pide calculado en el examen*

8. Economía de tokens a escala

Una empresa procesa 10 000 documentos legales al mes. Cada documento tiene 5 000 tokens de entrada y requiere un resumen de 200 tokens.

- (a) Calcula el total de tokens de entrada y de salida al mes.
- (b) Calcula el coste mensual con Claude Sonnet (3 \$/M entrada, 15 \$/M salida).
- (c) Calcula el coste mensual con Gemini 1.5 Pro (1,25 \$/M entrada, 5 \$/M salida).

- (d) Un empleado humano cobra 3 000 \$/mes y resume 50 documentos al día (22 días laborables, es decir 1 100 documentos/mes). ¿Cuántos empleados se necesitan para los 10 000 documentos? ¿Cuál es el coste humano mensual?
- (e) Calcula el coste por documento en cada caso (IA con Sonnet, IA con Gemini, humano) y el factor de ahorro de la IA frente al humano.