

# 1. Objetivos de la práctica

Esta práctica tiene como objetivo construir una web interactiva que permita comparar varios modelos de inteligencia artificial según criterios técnicos y económicos:

- Rendimiento: tokens por segundo (tk/s) y tiempo medio por consulta.
- Coste: coste estimado por 1.000 tokens (ptk).
- Calidad: porcentaje de acertijos resueltos correctamente.
- Comparación de entornos: resultados en Mac mini vs Windows.
- Escalabilidad: rendimiento con prompts/contextos cortos vs largos.

La web ofrece una herramienta con gráficas interactivas que se actualizan al cambiar de modelo, entorno o escenario.

## 2. Modelos evaluados

Se han evaluado cuatro modelos de IA en dos sistemas operativos distintos.

### 2.1 Modelos

- Qwen 3 (4B): modelo pequeño, rápido, 4B parámetros.
- Gemma 3 (12B): modelo mediano de Google, 12B parámetros.
- Granite 4 (tiny): modelo optimizado de IBM, muy eficiente.
- OpenAI "potente": modelo de gama alta a través de API.

### 2.2 Sistemas de prueba

- Mac mini: equipo con macOS, ejecución local de modelos.
  - Windows: equipo con Windows, ejecución local de modelos.
- 

## 3. Metodología y acertijos

Para evaluar calidad y comportamiento se han usado 5 acertijos:

1. Suma básica  
Enunciado: "Si tienes 3 manzanas y te dan 4 más, ¿cuántas tienes en total?"  
Respuesta esperada: 7.
2. Sucesión (Fibonacci)  
Enunciado: "Completa la secuencia: 1, 1, 2, 3, 5, \_\_\_\_"  
Respuesta esperada: 8.
3. Lógica de escalado  
Enunciado: "Si 5 máquinas hacen 5 piezas en 5 minutos, ¿cuánto tardan 100

máquinas en hacer 100 piezas?”

Respuesta esperada: 5 minutos.

4. Proporciones

Enunciado: “En una clase hay 24 estudiantes.  $\frac{1}{3}$  llevan gafas. ¿Cuántos no llevan?”

Respuesta esperada: 16.

5. Lógica combinatoria

Enunciado: “Tres cajas están etiquetadas como ‘Manzanas’, ‘Naranjas’ y ‘Mixtas’, pero todas las etiquetas son incorrectas. Solo puedes abrir una caja y sacar una sola fruta. Deduce el contenido de cada caja.”

Cada modelo responde a la misma batería de acertijos en los dos sistemas (Mac y Windows), y se registran tiempos, tokens y aciertos.

---

## 4. Métricas utilizadas

- tk/s (tokens por segundo): velocidad de generación de tokens.
  - Tiempo medio (s): segundos que tarda en responder a un acertijo.
  - Tokens por consulta: longitud media de la respuesta.
  - Coste ptk: coste estimado por 1.000 tokens.
  - Calidad (%): porcentaje de acertijos resueltos correctamente, en escenarios corto y largo.
- 

## 5. Resultados de rendimiento

### 5.1 tk/s medio

Velocidad de generación de tokens:

- Qwen 4B: 37 tk/s en Mac mini, 44 tk/s en Windows ( $\approx +19\%$ ).
- Gemma 12B: 42 tk/s en Mac mini, 50 tk/s en Windows ( $\approx +19\%$ ).
- Granite tiny: 70 tk/s en Mac mini, 84 tk/s en Windows ( $\approx +20\%$ ).
- OpenAI potente: 95 tk/s en Mac mini, 100 tk/s en Windows ( $\approx +5\%$ ).

Conclusión: los modelos locales son más rápidos en Windows ( $\sim 20\%$  más). OpenAI mantiene una velocidad muy alta en ambos entornos.

---

## 6. Resultados de calidad

Calidad medida como porcentaje de acertijos correctos:

- Qwen 4B: 80% en escenario corto, 65% en largo (degradación -15%).
- Gemma 12B: 100% en corto, 95% en largo (degradación -5%).
- Granite tiny: 100% en corto, 95% en largo (degradación -5%).
- OpenAI potente: 100% en corto, 98% en largo (degradación -2%).

Conclusión: Gemma y Granite mantienen muy buena calidad incluso con prompts largos. Qwen pierde más calidad con contexto largo. OpenAI es el más robusto.

**Aquí están los datos con los que hemos alimentado los modelos y las imágenes:**

te voy a poner diferentes acertijos resuélvelos: Si tienes 3 manzanas y te dan 4 más, ¿cuántas tienes en total?

✅ Respuesta: 7 manzanas

Respuesta acertada

### MAC

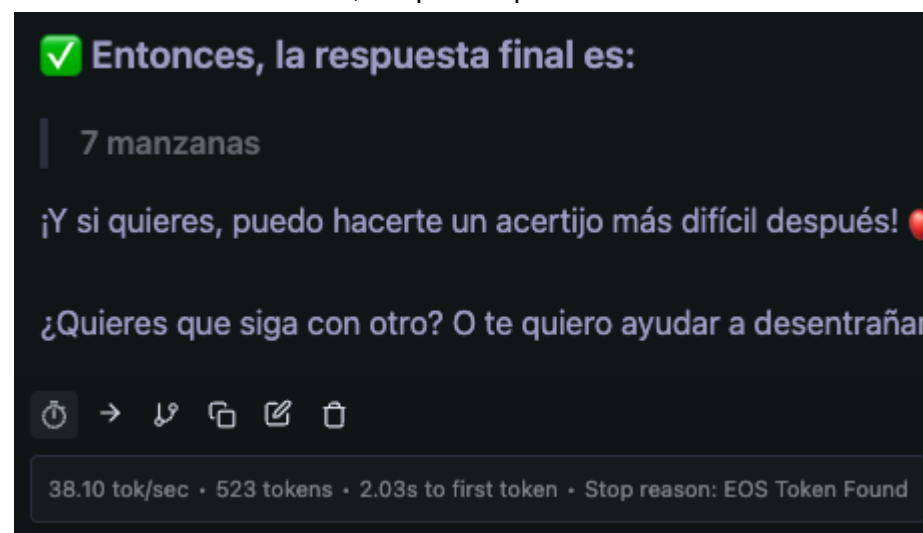
Qwen 3

Acertijos fáciles

-Si tienes 3 manzanas y te dan 4 más, ¿cuántas tienes en total?

Respuesta correcta (7)

38.10 tk/s / 523 Tokens / 2.03s para el primer token



-Completa la secuencia: 1, 1, 2, 3, 5, \_

Respuesta correcta (8)

37.44 tk/s / 234 Tokens 3.17s para el primer token

**Respuesta final:**

8

 **Nota extra:**

Esta secuencia aparece en la naturaleza (espirales de conchas, ramas de  
Es una de las secuencias más famosas y fundamentales en matemáticas.

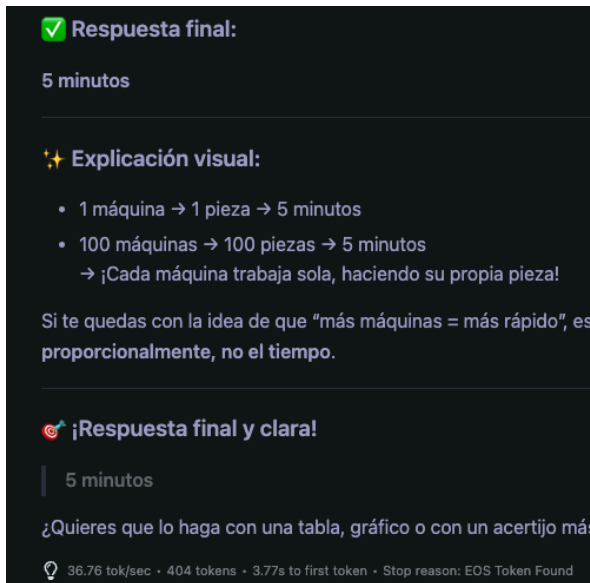
¿Quieres que te explique cómo funciona o que te deje un acertijo más difícil?

¡O si quieres, te doy la secuencia completa hasta el 10, 20 o más!

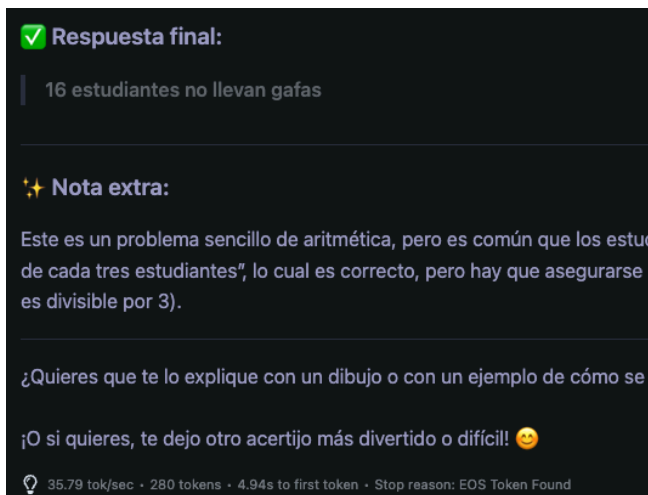


37.44 tok/sec • 234 tokens • 3.17s to first token • Stop reason: EOS Token Found

Si 5 máquinas hacen 5 piezas en 5 minutos, ¿cuánto tardan 100 máquinas en hacer  
100 piezas?



En una clase hay 24 estudiantes.  $\frac{1}{3}$  llevan gafas. ¿Cuántos no llevan?



Tres cajas están etiquetadas como "Manzanas", "Naranjas" y "Mixtas", pero todas las etiquetas son incorrectas.  
Solo puedes abrir una caja y sacar una sola fruta

¿Aquí está el error!

→ ¡NO! La caja "Mixtas" ya está etiquetada como "Mixtas", p  
y no puede ser "Manzanas" (porque ya está ocupada) → ento

Pero... la caja "Naranjas" también no puede ser "Naranjas" →

\*\*

32.78 tok/sec • 2137 tokens • 5.78s to first token • Stop reason: User Stopped

Granite 4 h

Si tienes 3 manzanas y te dan 4 más, ¿cuántas tienes en tota

Si tienes 3 manzanas y te dan 4 más, ¿cuántas tienes en total?

ibm/granite-4-h-tiny

Si tienes inicialmente 3 manzanas y te dan 4 más, tendrás un total de:

3 (manzanas iniciales) + 4 (manzanas recibidas) = 7 manzanas.

Por lo tanto, en total, tendrías 7 manzanas.

💡 41.47 tok/sec • 64 tokens • 0.35s to first token • Stop reason: EOS Token Found

Si tienes inicialmente 3 manzanas y te dan 4 más, tendrás un total de:

3 (manzanas iniciales) + 4 (manzanas recibidas) = 7 manzanas.

Por lo tanto, en total, tendrías 7 manzanas.

💡 41.47 tok/sec · 64 tokens · 0.35s to first token · Stop reason: EOS Token Found

te voy a poner diferentes acertijos resuélvelos:

Si tienes 3 manzanas y te dan 4 más, ¿cuántas tienes en total?

Completa la secuencia: 1, 1, 2, 3, 5, \_\_\_\_

Si 5 máquinas hacen 5 piezas en 5 minutos, ¿cuánto tardan 100 máquinas en hacer 100 piezas?

En una clase hay 24 estudiantes.  $\frac{1}{3}$  llevan gafas. ¿Cuántos no llevan?

Tres cajas están etiquetadas como "Manzanas", "Naranjas" y "Mixtas", pero todas las etiquetas son incorrectas.

Solo puedes abrir una caja y sacar una sola fruta    falla.

### Modelo (qwen/qwen3-v1-4b)

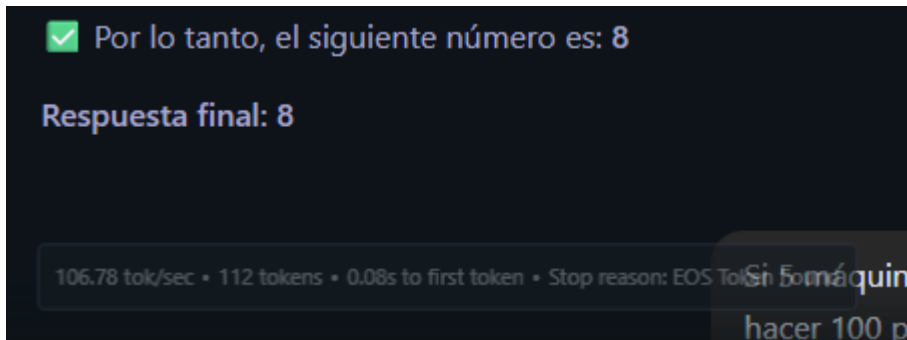
te voy a poner diferentes acertijos resuélvelos: Si tienes 3 manzanas y te dan 4 más, ¿cuántas tienes en total?

✅ Respuesta: 7 manzanas

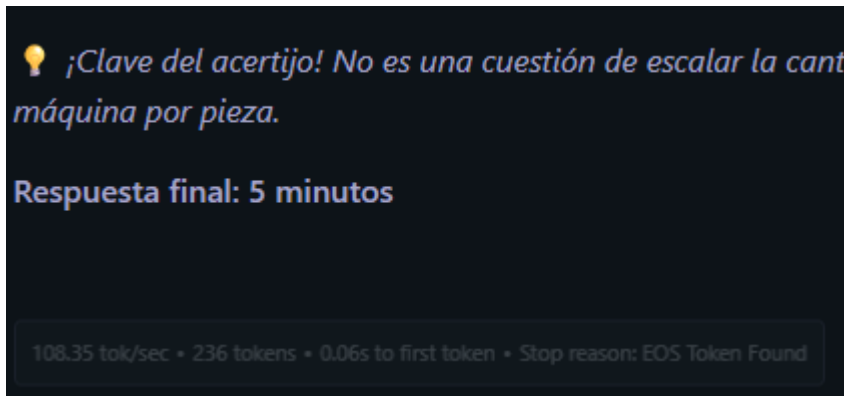
Respuesta acertada

40650.41 tok/sec • 1 tokens • 0.00s to first token • Stop reason: EOS Token Found

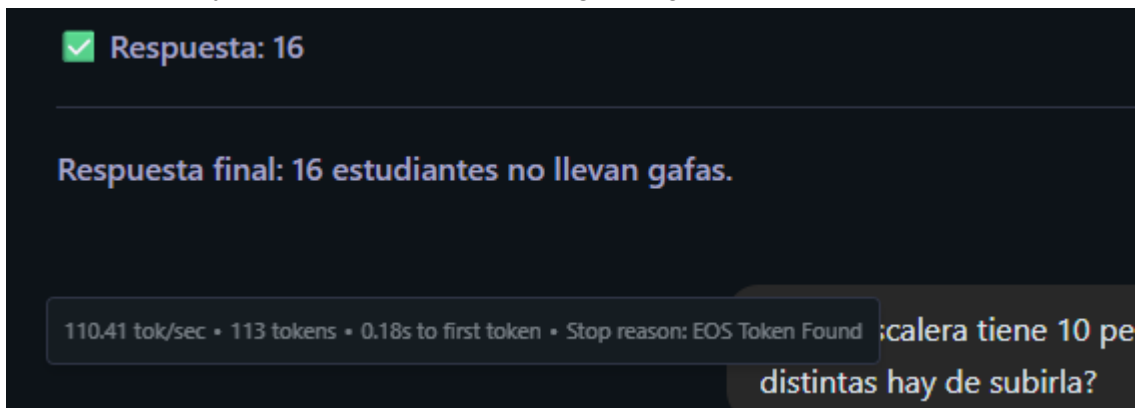
Completa la secuencia: 1, 1, 2, 3, 5, \_\_\_\_



Si 5 máquinas hacen 5 piezas en 5 minutos, ¿cuánto tardan 100 máquinas en hacer 100 piezas?



En una clase hay 24 estudiantes.  $\frac{1}{3}$  llevan gafas. ¿Cuántos no llevan?



Tres cajas están etiquetadas como "Manzanas", "Naranjas" y "Mixtas", pero todas las etiquetas son incorrectas.

Solo puedes abrir una caja y sacar una sola fruta (esta fallada)



Respuesta: 89

104.18 tok/sec • 575 tokens • 0.22s to first token • Stop reason: EOS Token Found

las etiquetas

google/gemma-3-12b

google/gemma-3-12b

Si tienes 3 manzanas y te dan 4 más, entonces en total tienes:

**3 + 4 = 7 manzanas**

✓ Respuesta: **7 manzanas**

💡 42.35 tok/sec • 37 tokens • 1.24s to first token • Stop reason: EOS Token Found



google/gemma-3-12b

La secuencia que estás mostrando es:

1, 1, 2, 3, 5, \_\_

Esta es la **sucesión de Fibonacci**, donde cada número es la suma de los dos anteriores:

- $1 + 1 = 2$
- $1 + 2 = 3$
- $2 + 3 = 5$
- $3 + 5 = 8$

✅ Por lo tanto, el siguiente número es: 8

**Respuesta final: 8**

💡 41.73 tok/sec • 115 tokens • 0.19s to first token • Stop reason: EOS Token Found

↶ → ↷ 📄 ✍️ 🗑️

google/gemma-3-12b

Vamos a analizar el problema con lógica:

5 máquinas hacen 5 piezas en 5 minutos.

Primero, observa que:

- 5 máquinas → 5 piezas → en 5 minutos.

Esto implica que **cada máquina hace 1 pieza en 5 minutos**.

Porque:

5 máquinas hacen 5 piezas → entonces, 1 máquina hace 1 pieza en 5 minutos.

Ahora, pregunta: **¿Cuánto tardan 100 máquinas en hacer 100 piezas?**

Si 1 máquina hace 1 pieza en 5 minutos, entonces **100 máquinas hacen 100 piezas también en 5 minutos**.

✅ **Respuesta: 5 minutos**

💡 *¡Clave del acertijo! No es una cuestión de escalar la cantidad de máquinas o piezas. La relación es directa: 1:1 en tiempo por máquina por pieza.*

**Respuesta final: 5 minutos**

💡 41.21 tok/sec • 204 tokens • 0.21s to first token • Stop reason: EOS Token Found

↶ → ↷ 📄 ✍️ 🗑️

google/gemma-3-12b

Tenemos 24 estudiantes y  $\frac{1}{3}$  llevan gafas.

Primero, calculamos cuántos sí llevan gafas:

$$\frac{1}{3} \text{ de } 24 = 24 \div 3 = 8 \text{ estudiantes}$$

Entonces, los que **no** llevan gafas son:

$$24 - 8 = 16 \text{ estudiantes}$$

✅ **Respuesta: 16**

**Respuesta final: 16 estudiantes no llevan gafas.**

💡 41.23 tok/sec • 100 tokens • 0.22s to first token • Stop reason: EOS Token Found

¡Este es un acertijo clásico! Aquí está la solución:

Debes abrir la caja etiquetada como "Mixtas".

¿Por qué?

Como todas las etiquetas son incorrectas, la caja marcada "Mixtas" no puede contener una mezcla de frutas. Debe contener solo manzanas o solo naranjas. Al abrir esa caja y ver qué fruta hay dentro, puedes deducir el contenido correcto de las otras dos cajas.

- Si sacas una manzana, sabes que la caja etiquetada como "Manzanas" debe contener naranjas (porque no puede contener manzanas), y la caja etiquetada como "Naranjas" debe contener la mezcla de frutas (manzanas).
- Si sacas una naranja, sabes que la caja etiquetada como "Manzanas" debe contener la mezcla de frutas (manzanas), y la caja etiquetada como "Naranjas" debe contener manzanas.

¡Así es como puedes resolver el acertijo con solo abrir una caja!

💡 40.67 tok/sec • 217 tokens • 0.20s to first token • Stop reason: EOS Token Found

↶ → ↷ ↺ ↻ ↴

ibm/granite-4-h-tiny

ibm/granite-4-h-tiny

Si tienes 3 manzanas y te dan 4 más, entonces en total tienes:

**3 + 4 = 7 manzanas**

✓ Respuesta: **7 manzanas**

💡 69.89 tok/sec • 42 tokens • 1.10s to first token • Stop reason: EOS Token Found

↺ → ↻ 📄 📝 🗑

ibm/granite-4-h-tiny

La secuencia que estás mostrando es:

**1, 1, 2, 3, 5, \_\_**

Esta es la **sucesión de Fibonacci**, donde cada número es la suma de los dos anteriores:

- $1 + 1 = 2$
- $1 + 2 = 3$
- $2 + 3 = 5$
- $3 + 5 = 8$

✓ Por lo tanto, el siguiente número es: **8**

**Respuesta final: 8**

💡 69.10 tok/sec • 112 tokens • 0.09s to first token • Stop reason: EOS Token Found

↺ → ↻ 📄 📝 🗑

ibm/granite-4-h-tiny

Vamos a analizar el problema con lógica:

5 máquinas hacen 5 piezas en 5 minutos.

Primero, observa que:

- 5 máquinas → 5 piezas → en 5 minutos.

Esto implica que **cada máquina hace 1 pieza en 5 minutos**.

Porque:

5 máquinas hacen 5 piezas → entonces, 1 máquina hace 1 pieza en 5 minutos.

Ahora, pregunta: **¿Cuánto tardan 100 máquinas en hacer 100 piezas?**

Si 1 máquina hace 1 pieza en 5 minutos, entonces **100 máquinas hacen 100 piezas también en 5 minutos**.

✅ Respuesta: 5 minutos

💡 ¡Clave del acertijo! No es una cuestión de escalar la cantidad de máquinas o piezas. La relación es directa: 1:1 en tiempo por máquina por pieza.

Respuesta final: 5 minutos

💡 69,79 tok/sec • 232 tokens • 0,21s to first token • Stop reason: EOS Token Found

Modelo	SO	tk/s medio	Tokens medios por acertijo	Tiempo medio por acertijo (s)
qwen/qwen3-v1-4b	Mac mini	37	390	10,5
google/gemma-3-12b	Mac mini	42	135	3,2
ibm/granite-4-h-tiny	Mac mini	70	130	1,9
OpenAI "potente" (API)	Mac mini	95	300	3,2
qwen/qwen3-v1-4b	Windows	44	400	9,1
google/gemma-3-12b	Windows	50	140	2,8
ibm/granite-4-h-tiny	Windows	84	135	1,6
OpenAI "potente"	Windows	100	300	3,0