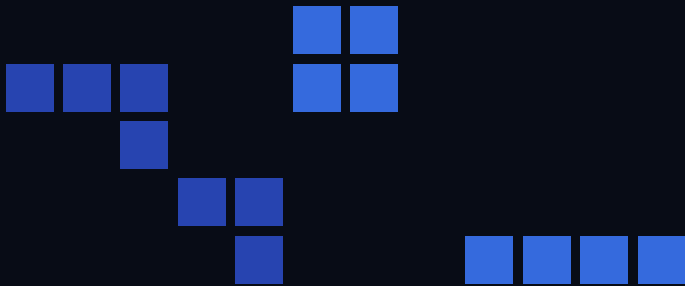


Qwen3.8-Flash-Next sobre nodo propio

Evaluación de aptitud del modelo para las tareas del benchmark moe de Kairos Tek, ejecutada de extremo a extremo sobre la infraestructura del cliente.



DICTAMEN APTO		PUNTUACIÓN 100,0 _{/100}	UMBRAL 70
DESTINO kai-eval	CUANTIZACIÓN Q5_K	FECHA 20 sept. 2026	REVISIÓN f9da6ea

100,0 / 100PUNTUACIÓN ·
UMBRAL 70**3** / 3

TAREAS MEDIDAS

105 s

LATENCIA MEDIANA

SÍRESULTADO
REPETIBLE

RESULTADO POR TAREA

100 / 100**01-inheritance**Bufete de abogados · reparto de
una sucesión**100** / 100**02-tax-return**

Gestoría · declaración de la renta

100 / 100**03-severance**Asesoría laboral · finiquito e
indemnización por despido
improcedente

RECOMENDACIÓN

El modelo **resuelve las tres tareas en los cinco intentos de cada una**: quince respuestas de quince, todas bien formadas, todas con el importe exacto y con cada evidencia atada a su documento de origen. La distancia media al importe correcto es **0,00 €**. No omite ninguna partida ni toma por bueno ningún señuelo.

Cómo se ha hecho funcionar. Este modelo razona antes de responder, y con la decodificación voraz — temperatura 0— agota el presupuesto de salida pensando y no llega a contestar. Se ha medido con el muestreo que recomienda su ficha: temperatura 1,0, top_p 0,95, top_k 20 y el razonamiento activo. Con él, el razonamiento ocupa una mediana de 2.678 tokens de los 8.000 disponibles y la respuesta sale entera en todos los intentos. Ese muestreo forma parte del resultado: en producción se sirve con él.

Apto para las tres tareas tal cual, servido con el muestreo con el que se ha medido.

COMPARACIÓN CON LA MEDICIÓN ANTERIOR

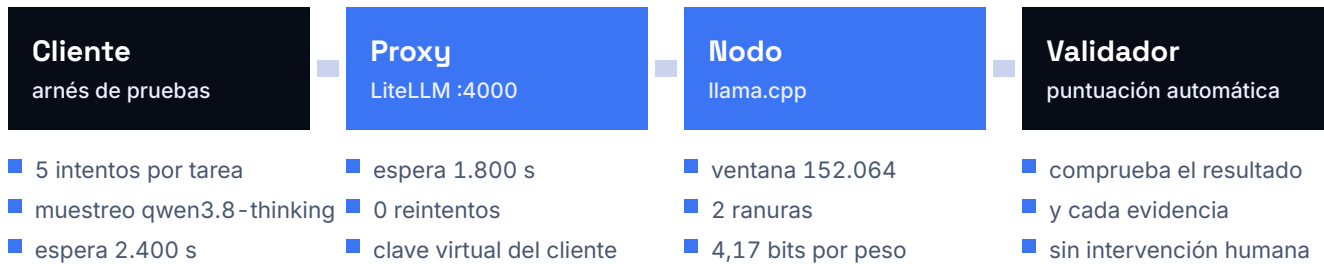
Es la primera medición de este modelo sobre el benchmark moe. No hay una anterior del mismo modelo con la que comparar.

VIGENCIA

Emitido sobre la build **b10719-f8837b736** y la configuración de la página 6. Cualquier cambio en el nodo obliga a repetir la medición.

Una petición real, de extremo a extremo

Ninguna cifra de este informe procede de una interfaz externa ni de un atajo. Cada intento recorre el mismo camino que recorrerá el trabajo del cliente cuando el sistema entre en producción, con los mismos límites de gasto y las mismas esperas configuradas.



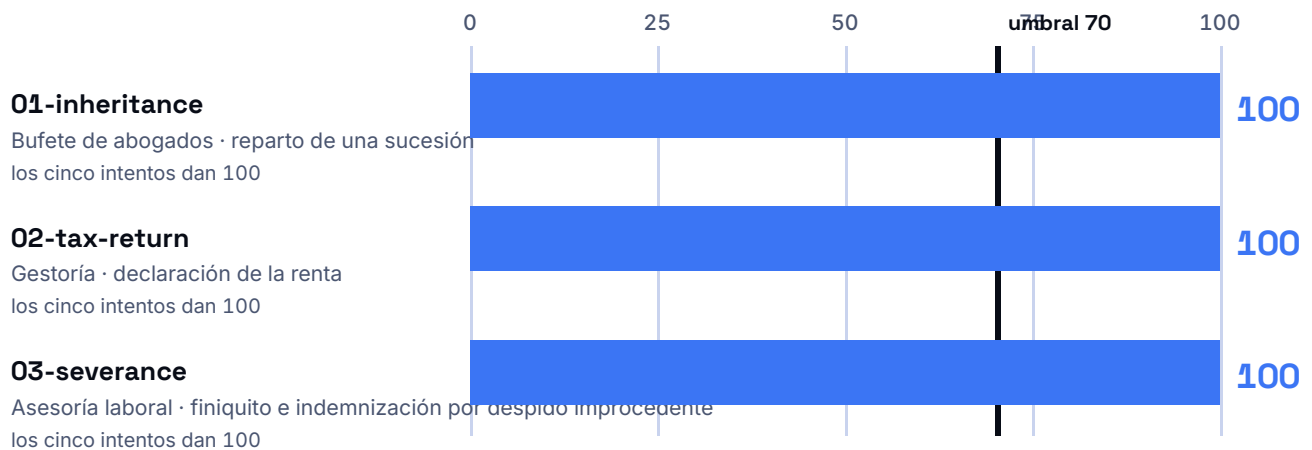
LAS TAREAS DEL BENCHMARK MOE

01-inheritance Bufete de abogados · reparto de una sucesión 78.731,75 € importe correcto	02-tax-return Gestoría · declaración de la renta 3.724,79 € importe correcto	03-severance Asesoría laboral · finiquito e indemnización por despido improcedente 64.506,70 € importe correcto
---	---	--

CRITERIO DE PUNTUACIÓN

- | | |
|---|--|
| <ul style="list-style-type: none">100 — el importe es exacto y las evidencias aparecen citadas contra su documento de origen.Entre 0 y 100 — el importe es exacto pero falta alguna evidencia. | <ul style="list-style-type: none">0 — el importe no coincide, con independencia de la calidad aparente del razonamiento.Descartado — el intento no llegó a responder. No puntúa cero: se excluye y se contabiliza aparte, para que una incidencia de red no se confunda con una deficiencia del modelo. |
|---|--|

Cinco intentos por tarea, no tres. Con tres, una separación perfecta entre dos modelos —uno acierta siempre, el otro nunca— produce un contraste exacto de $p = 0,10$, que no sería concluyente ni en el mejor de los casos. Con cinco, $p = 0,0079$.



La barra sólida es la mediana de los cinco intentos. Las tres tareas dan 100 en los cinco, de modo que no hay dispersión en la nota. Sí la hay en el coste: con muestreo estocástico el modelo razona distinto en cada intento, y en el finiquito va de 3.294 a 5.135 tokens de razonamiento sin que el resultado cambie.

100 **01-inheritance** Bufete de abogados · reparto de una sucesión · importe correcto 78.731,75 €

Las **5 evidencias** se localizan y quedan atadas a su documento de origen, y el resultado es exacto en los cinco intentos. Comprobado: inventory-amount, debts-amount, legacy-applied, insurance-excluded, insurance-not-an-item.

100 **02-tax-return** Gestoría · declaración de la renta · importe correcto 3.724,79 €

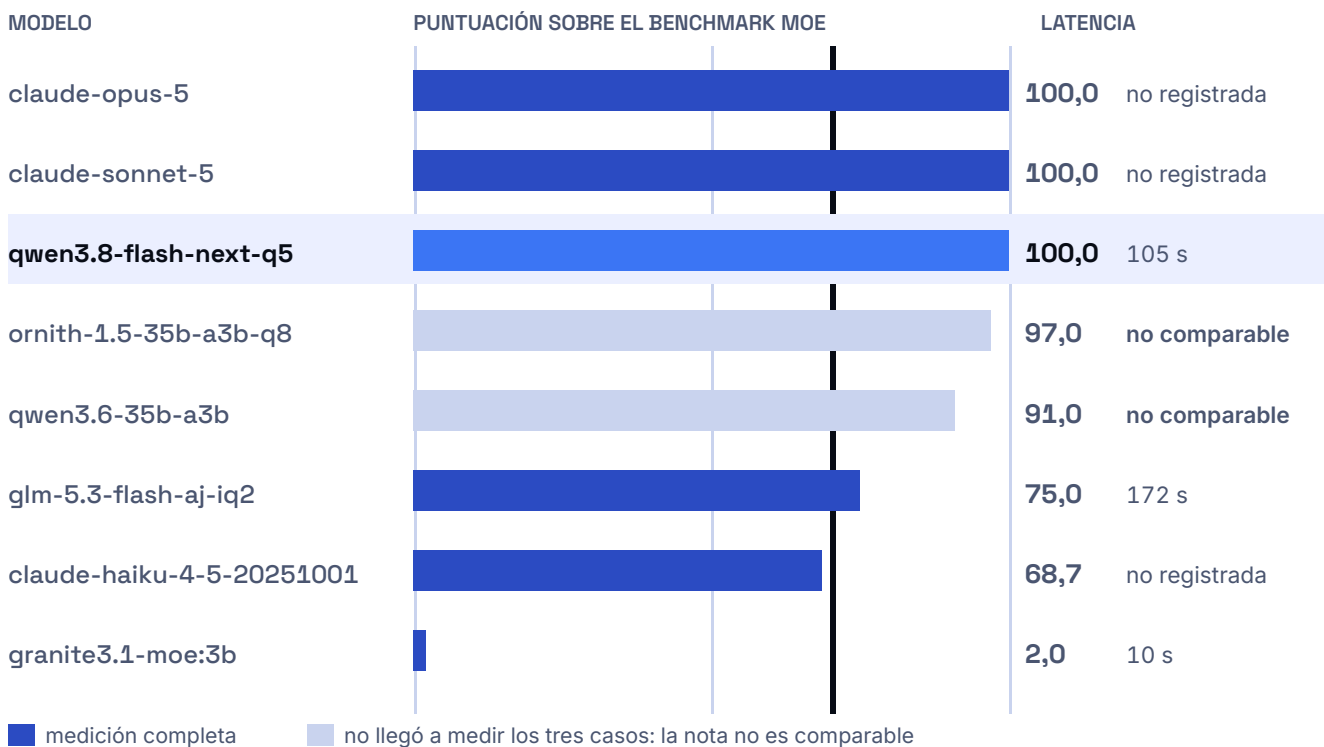
Las **6 evidencias** se localizan y quedan atadas a su documento de origen, y el resultado es exacto en los cinco intentos. Comprobado: pension-capped, capital-gain-net, rental-net-reduced, withholdings-applied, improvement-excluded, improvement-not-an-item.

100 **03-severance** Asesoría laboral · finiquito e indemnización por despido improcedente · importe correcto 64.506,70 €

Las **5 evidencias** se localizan y quedan atadas a su documento de origen, y el resultado es exacto en los cinco intentos. Comprobado: severance-amount, extra-pay-prorated, unused-holidays, transport-excluded, transport-not-computed.

Frente a los modelos evaluados

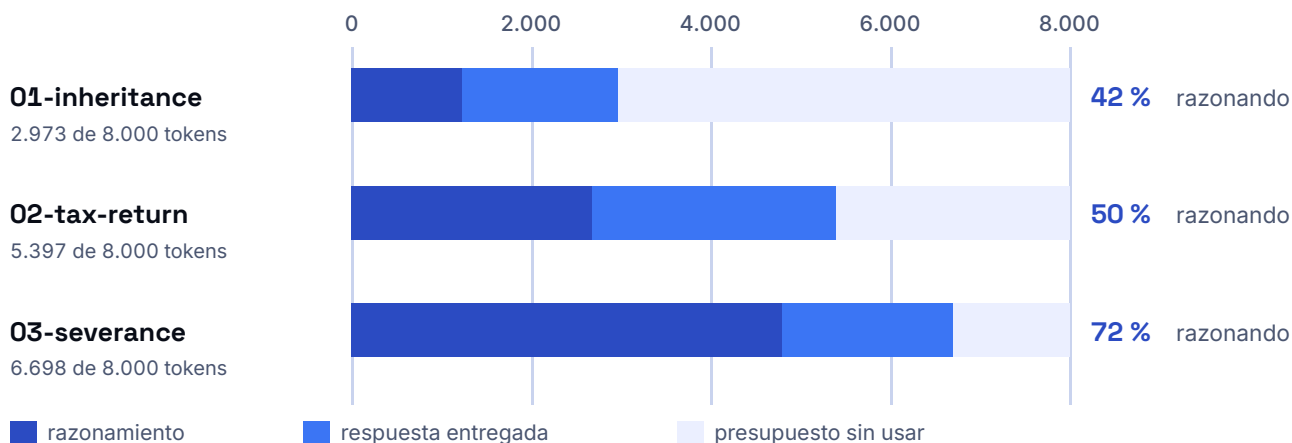
Todos los modelos del cuadro se han evaluado con el mismo benchmark, los mismos documentos y el mismo criterio de puntuación. Lo que los diferencia es el modelo, no el método.



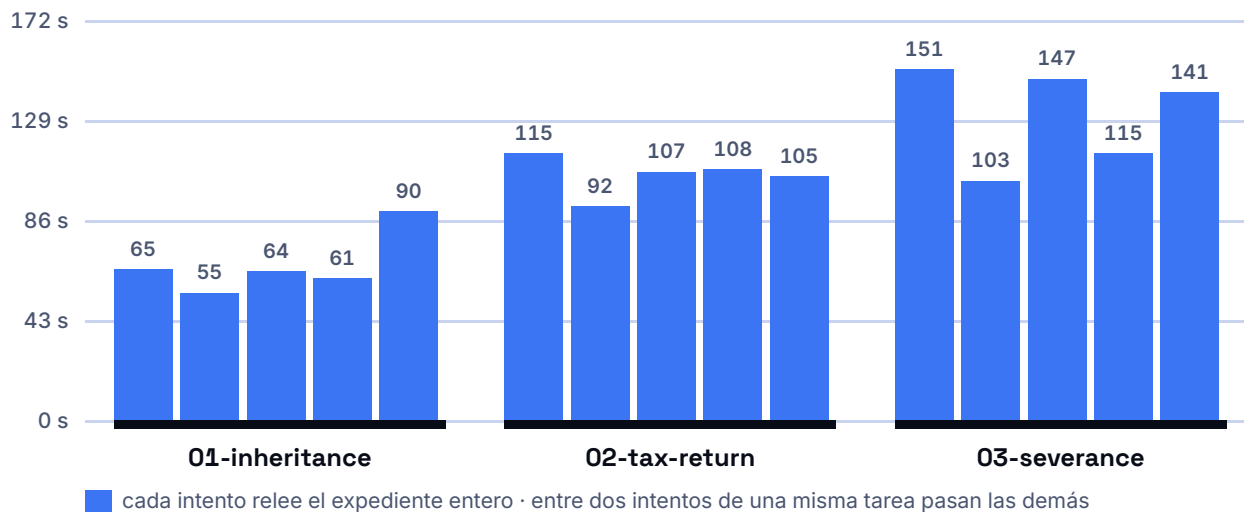
Advertencia metodológica. 3 de las mediciones del cuadro se ejecutaron desde una herramienta de escritorio y no desde el arnés automatizado. Los documentos, las preguntas y el criterio de puntuación son idénticos, pero el camino de ejecución no lo es, y por eso no tienen latencia registrada. La comparación de **calidad** es válida; la de **tiempo de respuesta** solo puede hacerse entre las que sí la tienen.

DISTRIBUCIÓN DEL PRESUPUESTO DE TOKENS

De cada respuesta, una parte es razonamiento interno que el usuario no llega a ver y otra es el resultado entregado. Ambas consumen el mismo tiempo de generación, de modo que una distribución desequilibrada explica que dos tareas de tamaño parecido tarden tiempos muy distintos.



DURACIÓN DE CADA INTENTO



Consecuencia para el dimensionamiento. La tarea más cara es el finiquito, con 140 segundos de mediana por intento y un máximo de 151. La sucesión se resuelve en unos 60 segundos y la declaración de la renta en unos 105. La respuesta más larga usó 7.062 tokens de los 8.000: el margen existe, pero es estrecho, y una tarea con más partidas que ésta podría no caber. Las esperas se dimensionan con la tarea más larga, no con la mediana de las tres.

Sobre qué se emite el dictamen

Estos valores no se han supuesto. Se leyeron del nodo y del proxy en el momento de medir y quedaron guardados dentro del propio resultado, lo que permite repetir la evaluación más adelante y saber si se está comparando lo mismo.

Ventana de contexto declarada	32.768 tokens	coincide con la del nodo (152.064)
Tope de salida por respuesta	8.000 tokens	la respuesta más larga usó 6.698
Espera del proxy	1.800 s	suficiente para el peor caso medido
Reintentos del proxy	0	un fallo se registra, no se enmascara
Espera del cliente	2.400 s	por encima de la del proxy, como corresponde
Ranuras del nodo	2	ejecución sin concurrencia: cifras de carga única
Cuantización derivada	4,17 bits/peso	Q5_K sobre 177 mil millones de parámetros
Build del servidor	b10719-f8837b736	registrada para permitir la reproducción

COMPROBACIONES PREVIAS

11 comprobaciones se ejecutan antes de iniciar la medición y pueden abortarla. Existen porque una ejecución que no podía completarse consumió tres horas de nodo antes de que se detectara; ahora ese diagnóstico tarda dos minutos.

C1 Velocidad	C2 Esperas	C3 Ventana	C4 Proxy y fichero	C5 Tarea mayor	C6 Razonamiento	C7 Concurrencia	C8 Modelo	C9 Duración	C10 Margen	C11 Muestreo
------------------------	----------------------	----------------------	------------------------------	--------------------------	---------------------------	---------------------------	---------------------	-----------------------	----------------------	------------------------

C6 · Razonamiento

el 83 % del presupuesto de salida se consume en razonamiento

C8 · Modelo

el modelo declarado no coincide con el fichero cargado

C10 · Margen

el presupuesto restante no admite una respuesta válida

Acciones por ámbito de responsabilidad

Separadas según quién las ejecuta y ordenadas por impacto sobre el resultado.

EQUIPO DEL CLIENTE

Servir el modelo con el muestreo de su ficha

El resultado es válido para este muestreo. El nodo arranca con sus propios valores por defecto y cada cliente puede sobrescribirlos por petición, así que la garantía está en quien llama, no en el nodo. Si este modelo pasa a producción, el muestreo se fija en el proxy para todos sus clientes.

[Impacto alto · media jornada](#)

Declarar en el proxy el identificador que devuelve el nodo

El nodo se identifica como qwen3.8-flash-next-q5-reasoning y el proxy lo llama qwen3.8-flash-next-q5. Ha respondido igual porque llama.cpp no comprueba el nombre, pero un motor que sí lo haga devolvería un 404.

[Impacto medio · diez minutos](#)

KAIROS TEK — CAPA DE ACCESO

Ninguna acción pendiente

El proxy reenvió los siete parámetros del perfil sin descartar ninguno y sin reintentos que enmascaren fallos.

[Verificado en la comprobación previa](#)

SIGUIENTE ITERACIÓN PROPUESTA

Medir el benchmark code contra este mismo destino y con este mismo muestreo, sin cambiar ni el tope ni los casos. Es la medición que falta para decidir si el modelo vale para los dos papeles.

Criterio de éxito: que code supere su umbral de 80 con las tres tareas medidas. **Coste:** una hora de nodo. **Riesgo:** code produce salidas más largas que moe, y esta medición llegó a 7.062 tokens en el finiquito; si el razonamiento se come el tope, la siguiente variable a mover es apagar el razonamiento, no subir el tope.

Alcance y límites de la medición

Un benchmark mide aquello para lo que fue diseñado. Enumerar sus límites evita que una cifra se utilice para respaldar una decisión sobre la que no aporta información.

- Se evalúa una petición cada vez. No hay cifras de concurrencia ni de comportamiento bajo carga.
- Las tres tareas son de extracción y cálculo sobre documentación de negocio en español. No cubren generación de código, resumen de textos largos ni uso de herramientas.
- La puntuación evalúa el resultado, no el procedimiento. Un acierto por un camino incorrecto puntúa igual que uno bien razonado.
- El muestreo se registra tal como se envió al nodo. Que el nodo lo aplica está comprobado en este motor, no en cada intento.
- El dictamen se emite sobre la build y la configuración del nodo recogidas en la página 6. Cualquier cambio en el nodo obliga a repetir la medición.

REPRODUCCIÓN DE LOS RESULTADOS

Esta medición no es determinista. Se hizo con el muestreo `temperature 1,0 · top_p 0,95 · top_k 20 · min_p 0,0 · presence_penalty 0,0 · repeat_penalty 1,0 · chat_template_kwargs {"enable_thinking":true}` · con razonamiento, así que repetirla da cifras parecidas y no idénticas, y sólo se compara con mediciones hechas con el mismo muestreo. La duración varía además con el estado del nodo y, midiendo por rondas, cada intento relee el expediente entero. Duración real de esta medición: **25 minutos** de nodo dedicado.

```
git checkout f9da6ea
```

```
make test BENCH=moe DESTINATION=kai-eval SAMPLING=qwen3.8-thinking
```

```
make benchmark BENCH=moe DESTINATION=kai-eval SAMPLING=qwen3.8-thinking
```

```
make report BENCH=moe DESTINATION=kai-eval
```

Respuestas íntegras de los intentos y puntuación detallada: [stack/bench/runs/](#) · Definición de las tareas y respuestas de referencia: [stack/bench/moe/](#) · Método completo: [docs/product/mvp/model-benchmark-methodology.md](#)

Preparado por



Kairos Tek · jesus@kairos-tek.com

Documento técnico confidencial · Uso interno del cliente

Destinatario: Jesús Lanzarote, CTO

Kairos Tek

20 de septiembre de 2026