Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Qué modelo de IA usar para cada tarea

Criterio de decisión para sistemas de comunicación reales.

Si llegaste acá es porque viste el short en el cual hablo sobre los mejores modelos en base a la tarea a realizar. Ahí dije que el error más caro no es elegir un modelo malo, es elegir **un solo modelo para todo el flujo**. Este documento es el desglose de eso. La idea no es dar un ranking de "el mejor modelo 2026", porque esa pregunta está mal hecha. Te voy a dar el criterio que uso para asignar un modelo a cada tarea dentro de un sistema que ya está corriendo en producción.

> **La idea base.** Un sistema de comunicación no es una sola tarea. Es una cadena: clasificar lo que entra, decidir qué hacer, generar una respuesta, a veces extraer datos, a veces resumir para pasarle el caso a una persona. Cada capa tiene una exigencia distinta. Pagás más cuando tratás a todos los eslabones como si fueran el más exigente.

El trade-off real de cada modelo

Esto es lo que importa para decidir, no los puntajes de un leaderboard. Lo ordené por la consecuencia que tiene en tu operación, no por capacidad técnica.

Tipo de modelo	Para qué es fuerte	El costo real que nadie te aclara
Razonamiento fuerte, costo medio (Ej. ChatGPT)	Decisiones con varios pasos, lógica encadenada, cuando el modelo tiene que "pensar" antes de responder.	Bueno como caballo de batalla, pero si lo usás para tareas triviales estás pagando capacidad que no se nota en el resultado final.
Contexto largo, mejor escritura (Ej. Claude)	Respuestas que el cliente va a leer, manejo de conversaciones largas, redacción que no suena a robot, generación de código.	Calidad alta, pero el costo por interacción suele ser bastante mayor. Justificable donde el cliente ve la salida. Caro si lo usás para clasificar.
Modelo liviano y barato (versiones mini / nano)	Volumen alto, tareas donde no necesitás una respuesta minuciosa, decisiones simples y rápidas.	El costo por mensaje es marginal. El riesgo es usarlo donde la calidad sí importa y degradar la experiencia sin darte cuenta.
Open source autoadministrado	Bajar el costo de inferencia a casi cero a escala, control total	El costo no desaparece, se muda. Pasa de la factura del proveedor a tu infraestructura:

Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Tipo de modelo	Para qué es fuerte	El costo real que nadie te aclara
(tipo Qwen y similares)	sobre los datos, sin enviar nada a un tercero.	GPUs, mantenimiento, monitoreo, y alguien que lo sostenga. Conviene a volumen y con equipo. No conviene para empezar.

> **Por qué la tabla comparativa te miente por omisión.** Las tablas de benchmark miden el modelo aislado, haciendo una sola cosa. Tu sistema no ejecuta una sola tarea. Un modelo que "gana" en razonamiento puede ser la peor decisión para clasificar diez mil mensajes por día, no porque sea peor, sino porque estás pagando 20 veces más por la misma salida que tendrías con modelos más económicos.

El criterio aplicado: tarea por tarea

Tomo un sistema de comunicación típico, uno de atención y ventas por WhatsApp con derivación a una persona cuando hace falta. Lo parto en sus tareas reales y asigno modelo a cada una. Este es el ejercicio que casi nadie hace, y es donde está la plata.

1. Clasificar la intención del mensaje que entra

MODELO LIVIANO Y BARATO

¿El cliente quiere precio, soporte, una queja, o agendar? Eso es una decisión de pocas categorías. No necesita razonamiento profundo ni redacción elegante. Es la tarea de mayor volumen del sistema y la que menos exige.

Acá es donde se quema más dinero sin que nadie lo mire: poner el modelo premium a clasificar es el caso de manual de costo disparado sin mejora de resultado. La clasificación con un modelo liviano bien instruido funciona igual de bien a una fracción del costo.

2. Extraer datos de un mensaje (nombre, mail, fecha, dirección)

MODELO LIVIANO Y BARATO

Sacar un email o una fecha de un texto desordenado, o de un audio ya transcripto, es una tarea acotada y verificable. La salida la podés validar con una regla, no necesitás el mejor modelo para esto.

Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Si el dato es crítico (un monto, un número de operación), el patrón correcto no es subir de modelo, es agregar una validación encima. Más modelo no te da más certeza acá. Una verificación sí.

3. Generar la respuesta que el cliente va a leer

CONTEXTO LARGO / MEJOR ESCRITURA

Esto es lo único que el cliente realmente percibe. Acá la calidad de redacción y el tono no son un lujo, son el producto. Una respuesta a un cliente enojado, una explicación clara, un mensaje que no suene a formulario.

Es el lugar donde el modelo más caro se justifica, porque la diferencia se ve y tiene consecuencia comercial directa. Pero ojo: se justifica acá, no en los pasos anteriores. Ese es el punto entero del documento.

4. Resumir la conversación antes de pasarla a una persona

CONTEXTO LARGO, SEGÚN EL CASO LIVIANO ALCANZA

Cuando el bot deriva a un agente humano, ese agente no debería leer treinta mensajes. Necesita un resumen útil. Si las conversaciones son largas y enredadas, contexto largo. Si son cortas, un modelo liviano resume bien y barato.

Regla práctica: la decisión del modelo acá depende del largo real de tus conversaciones, no de una preferencia general. Medilo antes de decidir.

5. Decisiones con varios pasos o lógica de negocio encadenada

RAZONAMIENTO FUERTE, COSTO MEDIO

“Si el cliente es de tal segmento y pide tal cosa fuera de horario, ofrecé esto y agendá para mañana.” Eso es razonamiento encadenado. Acá el modelo de razonamiento fuerte rinde, pero es la tarea menos frecuente del sistema, no la más.

El error clásico es elegir el modelo del sistema entero pensando en este caso, que aparece en el 5% de las interacciones, y pagarlo en el 100%.

Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

El número que ordena todo

Clasificar la intención de un mensaje y generar una respuesta empática a un cliente enojado no requieren lo mismo, lo dije en el video. Puesto en números aproximados para que se entienda la magnitud: la tarea de clasificación puede costar dos órdenes de magnitud menos que la de generación. Si las hacés con el mismo modelo premium, no estás comprando calidad en la clasificación, no se nota ahí. Estás regalando margen.

> **Cómo se ve esto en una factura real.** En un sistema con volumen, la clasificación y la extracción son la enorme mayoría de las llamadas al modelo. La generación de respuesta es la minoría. Si movés solo esas dos tareas de alto volumen a un modelo liviano, el resultado al cliente no cambia y la factura puede caer fuerte. Esa es la optimización con mejor relación esfuerzo/impacto que existe en estos sistemas, y casi nadie la hace porque eligieron un modelo el primer día y no lo volvieron a tocar.

Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Cuándo el open source tiene sentido (y cuándo no)

Los modelos open source autoadministrados te bajan el costo de inferencia a casi cero. Pero el costo no se evapora, se muda de lugar. Sale de la factura del proveedor y entra en tu infraestructura.

Tiene sentido cuando:

- Tenés volumen alto y sostenido. A poco volumen, el costo de mantener la infra supera lo que te ahorrás.
- Tenés un requisito real de que los datos no salgan a un tercero. Esto sí es una razón legítima de peso, no una excusa.
- Tenés a alguien que pueda sostener GPUs, actualizaciones y monitoreo. Si no lo tenés, el ahorro es ficticio.

No tiene sentido cuando:

- Estás empezando o validando. Acá la velocidad de iteración vale más que el ahorro de inferencia.
- Tu volumen es bajo o irregular. La cuenta no cierra.
- Estás eligiendo open source para “ahorrar” sin haber calculado el costo de operarlo. Eso no es ahorrar, es mover el problema y no mirarlo.

El resumen en una tabla de decisión

Si la tarea es...	Usá	Porque
Clasificar intención / enrutar	**Liviano y barato**	Alto volumen, baja exigencia
Extraer datos puntuales	**Liviano + validación**	Acotado y verificable
Respuesta que ve el cliente	**Contexto largo / escritura**	Es lo único que se percibe
Resumen para handoff humano	**Depende del largo real**	Medí tus conversaciones primero
Lógica encadenada de negocio	**Razonamiento, costo medio**	Poco frecuente, no pagues por todo
Volumen masivo + equipo + dato sensible	**Open source propio**	El ahorro recién cierra ahí

Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

LA PREGUNTA CORRECTA

No es “cuál modelo es mejor”. Es “cuál se adapta mejor a cada una de mis tareas”. El sistema no elige un modelo. Elige uno por tarea.

Si estás corriendo un sistema de comunicación y querés saber dónde específicamente estás pagando de más, eso es exactamente lo que reviso en una auditoría. No es una venta acá, es el siguiente paso lógico si lo de arriba te hizo ruido sobre tu propio sistema.

Cualquier duda concreta sobre tu caso, escribime a gonzalo@ggomez.dev y analizamos el caso en concreto.