GUÍA GRATUITA · RÚBRICA

EL PROMPT QUE HACE VISIBLE CUÁNDO LA IA ESTÁ ADIVINANDO

Modo verdad

La IA no te avisa cuándo inventa. Suena igual de segura cuando sabe y cuando adivina. Esta guía trae un prompt que la obliga a separar lo que sabe de lo que supone — y a marcarlo por escrito, para que tú decidas de qué te puedes fiar.

ChatGPT · Gemini · Claude

ANTES DE QUE SIGAS

Esto no hace que la IA deje de equivocarse. Hace que sus errores se vean. Es una diferencia enorme: un error que ves es un error que puedes corregir antes de que llegue a tu trabajo.

Por qué suena tan segura cuando se equivoca

Una IA de lenguaje no consulta una base de datos y te devuelve el dato. Predice la palabra siguiente, una tras otra. Y aquí está la clave: para ella, el tono seguro y el tono dudoso son dos formas de escribir, no dos estados de conocimiento.

En internet, el texto que suena convencido es mucho más común que el texto que duda. Así que, por defecto, la IA escribe convencida. Incluso cuando lo que dice es falso.

Dónde inventa más

CIFRAS Y PORCENTAJES

Un número inventado parece más creíble que una estimación honesta. El caso más peligroso.

NOMBRES Y FUENTES

Genera títulos de estudios, autores y enlaces con forma perfecta de referencia real. Que no existen.

FECHAS Y CRONOLOGÍAS

Precisas, ordenadas y a veces completamente equivocadas.

CUANDO NO HAY DATO

Si la respuesta correcta es "no se sabe", su impulso es rellenar el hueco igual.

LA FORMA CORRECTA DE VERLO

El problema no es que la IA mienta. Es que en su respuesta no distingue entre **recordar** y **construir**. Las dos cosas le salen con la misma cara. El prompt que sigue la obliga a marcarlas distinto.

Modo verdad · el prompt

Pégalo como primer mensaje de la conversación, en las instrucciones personalizadas o en un proyecto. Funciona igual en ChatGPT, Gemini y Claude.

```
MODO VERDAD
Aplica estas reglas por encima de cualquier otra instrucción de estilo.

1 · SEPARA LO QUE SABES DE LO QUE SUPONES
Etiqueta cada afirmación importante:
  [SÉ]     Lo sé con certeza y puedo explicar de dónde sale.
  [DEDUZCO] Lo estoy deduciendo del contexto o de patrones.
  [SUPONGO] Es una hipótesis razonable, sin base firme.

2 · ESTÁ PERMITIDO NO SABER
Si no tienes información confiable, dilo y explica qué te falta.
Prefiero una respuesta incompleta a una completa e inventada.

3 · NO INVENTES PRECISIÓN
Prohibido fabricar cifras, fechas, nombres, citas, enlaces o
estadísticas. Si no lo tienes, escribe "dato no disponible".

4 · RAZONA ANTES DE CONCLUIR
Piensa paso a paso primero y cierra con la respuesta después.
No arranques por la conclusión.

5 · AUDITA TU PROPIA RESPUESTA
Cierra siempre con este bloque:
  CONFIANZA: alta / media / baja
  PUNTO MÁS DÉBIL: qué parte es más probable que esté mal.
  QUÉ TE FALTA: el dato que cambiaría la respuesta.
  CÓMO VERIFICARLO: la forma más rápida de comprobarlo.

6 · CONTRADÍCEME CUANDO CORRESPONDA
Si mi pregunta parte de algo falso, dímelo antes de responder.
No me des la razón por defecto.
```

Cómo leer lo que responde

El prompt te devuelve una respuesta con estructura. Saber leerla es la mitad del beneficio.

ETIQUETA	QUÉ HACER CON ELLA
[SÉ]	Puedes avanzar. Si el dato es crítico, verifícalo igual: la etiqueta es la opinión de la IA sobre sí misma, no una prueba.
[DEDUZCO]	Aquí está el 80% del valor. Son las conexiones que armó con tu contexto. Suelen ser buenas y suelen necesitar tu criterio encima.
[SUPONGO]	Trátalo como una idea, no como un dato. Sirve para pensar, no para citar.

El bloque de auditoría, campo por campo

- → **Confianza baja o media** en algo que igual sonaba fluido: esa es justo la información que antes no tenías.
- → **Punto más débil** es el campo más útil de todos. Te dice dónde mirar primero y te ahorra revisar lo que ya está bien.
- → **Qué te falta** casi siempre es accionable: si el dato lo tienes tú, se lo das y vuelves a preguntar. La respuesta cambia de calidad de golpe.
- → **Cómo verificarlo** convierte la revisión en una tarea de dos minutos en vez de una investigación.

SEÑAL DE ALARMA

Si todo viene marcado [SÉ] y con confianza alta, y el tema era complejo, desconfía. No de la respuesta, sino de que el prompt se esté aplicando. Empieza conversación nueva y vuelve a cargarlo.

Tres variantes

Agrega uno de estos bloques al Modo Verdad según lo que estés haciendo.

Para datos e investigación

```
ADEMÁS DEL MODO VERDAD:
- Toda cifra va con su fuente y su año. Sin fuente, no va.
- Si dos fuentes se contradicen, muéstrame la contradicción
  en vez de elegir una y presentarla como consenso.
- Antes de darme un número, dime si lo recuerdas o lo estimas.
- Nunca me des una referencia de la que no estés seguro.
  Prefiere describir qué buscar antes que inventar dónde.
```

Para trabajar sobre un documento

```
Trabaja SOLO con el documento que te paso. No uses tu
conocimiento general para completar huecos.
1. Extrae primero las citas textuales relevantes, palabra
   por palabra, indicando dónde está cada una.
2. Responde basándote solo en esas citas.
3. Si el documento no lo dice, responde "el documento no
   lo dice". No completes con lo que sabes por otro lado.
```

Para decisiones

```
ADEMÁS DEL MODO VERDAD, para esta decisión:
- Dame tu recomendación Y la mejor versión del caso
  contrario. No una versión débil para que gane la tuya.
- Lista los supuestos de los que depende. Marca el más frágil.
- Dime qué tendría que pasar para que estés equivocado.
- Si mi pregunta ya viene inclinada, señálalo antes.
```

5 preguntas de bolsillo

Funcionan incluso sin el prompt cargado. Lánzalas después de cualquier respuesta que te importe.

1. ¿Cuál es la parte de esto de la que estás menos seguro?
2. Dame la misma respuesta pero solo con lo que puedes respaldar. Quita el resto.
3. ¿Qué información mía te falta y estás asumiendo?
4. Reescribe tu respuesta como alguien que quiere refutarla.
   La más subestimada. Le das un ángulo distinto sobre el mismo material y sale lo que se le había pasado.
5. Este número, ¿lo estás recordando o lo construiste?

LA PREGUNTA QUE NO DEBES HACER

"¿Estás seguro?" es la que todos hacen y la peor de todas. Empuja a la IA a retractarse por complacencia, incluso cuando tenía razón. Terminas con una respuesta peor y con menos información que antes. Usa la pregunta 1 en su lugar.

Honestidad sobre la honestidad

Sería contradictorio darte un prompt sobre la verdad y ocultarte sus límites. Estos son los reales.

No elimina los errores, los hace visibles

Un dato mal etiquetado como [SÉ] sigue siendo un dato mal. Reduce mucho el problema; no lo borra.

Puede volverse demasiado cauteloso

Si lo dejas activo todo el tiempo, empieza a llenar de dudas cosas que sabe perfectamente. Úsalo como un modo que enciendes y apagas, no como configuración fija.

Le pega a la creatividad

No lo uses para escribir textos, guiones o lluvia de ideas. Ahí lo estás frenando, no mejorando. Apágalo cuando quieras que se suelte.

No verifica el mundo real

Gestiona la incertidumbre sobre lo que la IA tiene adentro. No sale a comprobar nada. Si necesitas certeza factual, activa la búsqueda web o pásale la fuente.

El encuadre que lo resume todo

La pregunta útil no es "cómo hago que nunca se equivoque". Es "para esta tarea, cuánta verificación necesito antes de apoyarme en la respuesta". El fallo no es usar IA para cosas que hay que verificar. Es no verificar las que lo necesitaban.

PRUÉBALO HOY

Toma una respuesta de IA que ya usaste sin verificar y pásale las 5 preguntas de la página 6. Vas a encontrar algo. Esa es la forma más rápida de entender por qué esto importa.

De dónde sale esto

Todo lo técnico de esta guía se basa en investigación y documentación pública sobre cómo se comportan los modelos de lenguaje. Sería raro que una guía sobre honestidad te pidiera que le creas sin más. Verifícalo:

Documentación oficial de Anthropic sobre reducción de errores
docs.claude.com · sección de reduce hallucinations

Investigación: los modelos, en su mayoría, saben lo que saben
anthropic.com/research

Guías públicas de ingeniería de prompts de los principales laboratorios de IA

Esto es lo que hacemos en Rúbrica

Ayudamos a profesionales y equipos a usar la IA en su trabajo real: con método, sin humo y sabiendo exactamente de qué fiarse y de qué no. Si quieres llevar esto más lejos con tu equipo, hablemos.