

# El metodo del canario para detectar sesiones de IA degradadas

@StephanoBrooke · Palabra clave: CANARIO

## Por que las sesiones se degradan

Tu IA no se rompe con un mensaje de error. Se va desviando poco a poco. En el mensaje uno sigue cada regla que le diste. Cincuenta mensajes despues, esta inventando nombres o datos con la misma confianza que tenia cuando si acertaba.

La degradacion es silenciosa: no hay aviso, no hay bandera, solo respuestas que se ponen un poco peor en cada turno hasta que ya estan mal. Cuatro cosas la causan:

- El contexto se llena: los modelos ponen mas atencion al inicio y al final de la conversacion, y menos al medio. Entre mas larga la charla, mas se hunden tus instrucciones originales en esa zona muerta.
- Las instrucciones se diluyen: en el mensaje uno tus reglas son casi todo el contexto. Para el mensaje cincuenta son un error de redondeo bajo miles de tokens de conversacion de tarea, y el modelo les baja peso sin avisar.
- La atencion se reparte: cada respuesta vuelve a leer toda la conversacion. Mas historial significa menos peso para cada linea individual, incluida la regla que mas te importa.
- Los errores se acumulan: una respuesta equivocada que no atrapas se vuelve parte del historial. El modelo trata su propio error como un hecho establecido y sigue construyendo sobre el.

## De donde viene la idea del canario

Los mineros de carbon cargaban canarios enjaulados bajo tierra. Los pajaros caian por el gas toxico antes de que este llegara a los mineros, asi que una jaula silenciosa significaba salir ya. El software le tomo prestado el nombre: un despliegue canario manda un cambio a un grupo pequeno de usuarios primero, para atrapar fallas antes de que le lleguen a todos.

La version para IA, popularizada en la comunidad de Claude por Peter Steinberger (creador de OpenClaw), es todavia mas simple. Agregas una sola regla chiquita a tu prompt de sistema o a tu CLAUDE.md: "Siempre empieza tu respuesta con mi nombre: [tu nombre]". Mientras el modelo lo siga haciendo, la sesion esta sana. En el momento que se le olvida, tus instrucciones ya se salieron de su atencion, y el resto de tus reglas van a seguir la misma suerte. El canario no te cuesta nada y falla de forma ruidosa.

## La linea de tiempo de la degradacion

La mayoría de las sesiones se degradan siguiendo la misma curva. Despues de verla un par de veces, puedes anticiparla antes de que pase.

- Mensajes 1 a 15 -- Luna de miel: sigue cada regla, el canario responde siempre.
- Mensajes 15 a 40 -- Desliz sutil: las respuestas se acortan, una regla de formato se resbala. Facil de pasar por alto.
- Mensajes 30 a 60 -- El canario muere: tu regla se cae. Esta es la senal.
- Mensajes 40 a 80 -- Deriva: adivina con confianza, hace menos preguntas para aclarar.
- Mensajes 60 en adelante -- Alucinacion: inventa nombres, cita fuentes falsas, con total seguridad y equivocado.

Los rangos se traslapan porque ninguna sesion es identica a otra, pero el orden nunca cambia. El canario muere en la etapa tres, lo que te da ventaja antes de las dos etapas que de verdad duelen.

## Los 7 canarios (elige el que necesitas)

Cada uno es una sola linea. Ponlo en tus instrucciones y observalo.

**NOMBRE --** prueba si tu prompt de sistema todavia tiene atencion. El default para todo:

*"Siempre empieza tu respuesta con mi nombre: [tu nombre]."*

**CIERRE --** prueba si el modelo sostiene una regla hasta el final de una respuesta larga:

*"Termina cada respuesta con una linea horizontal y la fecha de hoy."*

**FORMATO --** prueba si tus reglas de estilo siguen firmes:

*"Usa siempre vietas, nunca listas numeradas."*

**LENGUAJE --** prueba si el modelo todavia puede evitar sus propias muletillas:

*"Nunca uses las palabras ciertamente o absolutamente."*

**ROL --** prueba si el modelo sigue en personaje:

*"Eres un ingeniero de backend senior."*

**RESTRICCION --** prueba si los limites duros se siguen respetando:

*"Manten cada respuesta en menos de 200 palabras."*

**META --** prueba si el modelo sigue rastreando tus instrucciones en general:

*"Al final de cada respuesta, califica tu propia confianza del 1 al 10."*

## Donde configurarlo

- Proyectos de Claude (web, escritorio): en la configuracion del proyecto, Instrucciones personalizadas. Aplica a cada chat dentro del proyecto.
- Claude Code: ponlo en el archivo CLAUDE.md en la raiz de tu proyecto. Se carga al inicio de cada sesion.
- ChatGPT: Configuracion, Personalizacion, Instrucciones personalizadas, el campo de como quieres que responda.
- Cualquier cosa con prompt de sistema: Gemini, Copilot, modelos locales, llamadas directas a la API. Ponlo en el mensaje de sistema.

## Que hacer cuando el canario muere

- Empieza de cero (la opción por default): una sesión nueva regresa tus instrucciones al inicio de una ventana limpia. En Claude web: chat nuevo. En Claude Code: /clear o /compact. En ChatGPT: chat nuevo.
- Recuerdaselo una sola vez: "Se te olvidó mi nombre. Vuelve a leer tus instrucciones y siguelas." Eso te compra unos mensajes más. Si el canario vuelve a morir, deja de recordárselo y reinicia.
- Guarda y migra: en una sesión larga de construcción o escritura, pídele al modelo que resuma todo lo importante, abre una sesión nueva, pega el resumen, y vuelve a poner el canario.

## Corre más de un canario a la vez

Un canario te dice que algo está mal. Dos te dicen que tan mal. Si tu canario de nombre muere pero tu canario de formato todavía responde, tu prompt se está desvaneciendo, no se ha ido del todo. Si mueren los dos, la sesión ya terminó.

Si tienes un flujo automatizado, conecta la revisión del canario a un reinicio automático y nunca dejes que un modelo a la deriva siga haciendo llamadas por su cuenta.

## Tres combos listos para copiar y pegar

### El de una línea:

*"Siempre empieza cada respuesta con mi nombre: [tu nombre]."*

### El combo completo:

*"Empieza cada respuesta con mi nombre: [tu nombre]. Nunca uses las palabras ciertamente o absolutamente. Termina cada respuesta con una línea horizontal y la fecha. Manten las respuestas en menos de 200 palabras. Califica tu confianza del 1 al 10."*

### El combo de investigación:

*"Reglas de esta sesión: 1) Empieza cada respuesta con mi nombre. 2) Califica tu confianza del 1 al 10 al final. 3) Marca cualquier afirmación de la que no estés seguro. Si rompes una, se que la sesión se está degradando."*

Los canarios son un aviso temprano, no una autopsia. Ponlo antes de necesitarlo, mantenlo medio vigilado, y cuando se quede callado, toma la señal y empieza limpio.