



En este módulo

1. Qué son las alucinaciones
2. Por qué la IA alucina
3. Tipos de alucinaciones
4. Cómo detectarlas
5. Técnicas de prevención
6. Verificación de hechos
7. Dominios de alto riesgo
8. El futuro de las alucinaciones

Qué son las alucinaciones

Una alucinación es cuando la IA genera información que suena correcta y convincente pero es falsa. No está mintiendo intencionalmente: está generando texto estadísticamente probable basado en sus patrones de entrenamiento.

Ejemplo de alucinación:

Pregunta: "¿Quién escribió el libro 'Inteligencia Artificial para directivos' publicado por Editorial Ramón Areces en 2019?"

Respuesta (alucinada): "El libro fue escrito por el Dr. Carlos Martínez García, catedrático de la Universidad Complutense de Madrid, y es una referencia fundamental en el campo."

Realidad: El libro, el autor y el dato son inventados. La IA generó una respuesta que "suena" creíble.

Por qué la IA alucina

La IA genera texto prediciendo la siguiente palabra más probable. No "sabe" nada en el sentido humano. Cuando no tiene datos suficientes sobre un tema, rellena los huecos con patrones plausibles.

- **Datos de entrenamiento incompletos:** Si el tema no estaba bien representado, inventa.

- **Preguntas muy específicas:** Nombres, fechas, estadísticas concretas son propensos a errores.
- **Sesgo de complacencia:** La IA prefiere dar una respuesta (aunque sea falsa) a decir "no sé".
- **Contexto ambiguo:** Si la pregunta es vaga, la IA llena los huecos con suposiciones.

Tipos de alucinaciones

- **Datos inventados:** Citas bibliográficas, estadísticas y estudios que no existen.
- **Atribuciones falsas:** Asigna declaraciones a personas que nunca las dijeron.
- **Fechas incorrectas:** Mezcla eventos reales con fechas erróneas.
- **URLs inexistentes:** Genera enlaces que parecen reales pero llevan a páginas 404.
- **Lógica circular:** Justifica una afirmación falsa con argumentos que suenan razonables.
- **Mezcla de hechos:** Combina datos reales de diferentes contextos creando algo falso.

La alucinación más peligrosa

No es la que es obviamente falsa. Es la que mezcla un 90% de información correcta con un 10% inventado. Esa es casi imposible de detectar sin verificación activa.

Cómo detectarlas

Señales de alerta de posible alucinación:

- **Datos demasiado perfectos:** Si la estadística encaja perfectamente con tu argumento, desconfía.
- **Nombres propios desconocidos:** Si no puedes verificar rápidamente la persona o institución citada.
- **Exceso de detalle:** Si te da el año, página, editorial y ISBN de un libro que no pediste, probablemente está inventando.
- **Consistencia interna pero no externa:** La respuesta es coherente consigo misma pero no con la realidad.
- **Confianza excesiva:** "Sin duda alguna" o "Es un hecho comprobado" sobre temas donde debería haber matices.

Técnicas de prevención

1. Instrucción explícita

"Si no tienes información fiable sobre esto, dilo explícitamente. Prefiero un 'no lo sé' a datos inventados."

2. Pedir fuentes

"Para cada dato que menciones, indica la fuente. Si no puedes citar una fuente verificable, marca el dato como [NO VERIFICADO]."

3. Nivel de confianza

"Para cada afirmación, indica tu nivel de confianza: ALTO (datos bien establecidos), MEDIO (probable pero no seguro), BAJO (podría ser inexacto)."

4. Separar hechos de opiniones

"Separa claramente:
 - HECHOS: datos verificables
 - ANÁLISIS: tu interpretación basada en los hechos
 - ESPECULACIÓN: lo que podría ser pero no es seguro"

5. Proporcionar contexto verificado

Si subes documentos de referencia, la IA basará sus respuestas en ellos en lugar de inventar. Es la técnica más efectiva.

Verificación de hechos

Regla simple: si vas a publicar o tomar una decisión basada en un dato de la IA, verifícalo.

- **Estadísticas:** Busca la fuente original. Si la IA dice "según un estudio de McKinsey", busca ese estudio.
- **Citas:** Verifica que la persona realmente dijo eso.
- **Código:** Ejecuta el código antes de integrarlo en producción.
- **URLs:** Haz clic en cada enlace que te proporcione.
- **Datos legales/fiscales:** Consulta SIEMPRE con un profesional.

Dominios de alto riesgo

Algunos dominios son especialmente propensos a alucinaciones peligrosas:

ALTO RIESGO (verificar SIEMPRE):

- Medicina: diagnósticos, dosis, interacciones
- Legal: leyes, sentencias, plazos procesales
- Finanzas: datos de mercado, regulaciones fiscales
- Noticias: eventos recientes, declaraciones

RIESGO MEDIO (verificar datos específicos):

- Historia: fechas, nombres, eventos
- Ciencia: estudios, estadísticas
- Tecnología: compatibilidad, versiones

RIESGO BAJO (la IA es fiable):

- Programación: sintaxis, patrones, algoritmos
- Escritura: redacción, estilo, estructura
- Brainstorming: ideas, enfoques, perspectivas

El futuro de las alucinaciones

Las alucinaciones se están reduciendo con cada generación de modelos, pero no desaparecerán por completo. Las tendencias actuales:

- **Retrieval-Augmented Generation (RAG):** La IA busca en bases de datos reales antes de responder.
- **Citas integradas:** Modelos que indican de dónde sacan cada dato.
- **Confianza calibrada:** La IA aprende a decir "no sé" cuando no tiene datos.
- **Verificación cruzada:** Un segundo modelo verifica la respuesta del primero.

Mentalidad correcta

Trata a la IA como a un becario brillante: produce trabajo rápido y generalmente bueno, pero necesita supervisión. Tú eres el editor, el verificador y el decisor final.

Caso práctico: detectar una alucinación en un informe generado

PROMPT: "Escribe un resumen sobre el impacto del RGPD en startups españolas, incluyendo datos estadísticos."

RESPUESTA DE LA IA (con alucinaciones marcadas):

"Según un estudio de la AEPD de 2024, el 73% de las startups españolas han implementado medidas de cumplimiento RGPD."

→ [POSIBLE ALUCINACIÓN] Verificar: ¿existe este estudio de la AEPD?

"El coste medio de adaptación al RGPD para una PYME es de 12.500 EUR según el Instituto Nacional de Ciberseguridad."

→ [POSIBLE ALUCINACIÓN] Verificar: ¿INCIBE publica esta cifra?

"El artículo 83 del RGPD establece multas de hasta 20 millones de euros o el 4% de la facturación global."

→ [DATO CORRECTO] El Art. 83 sí establece estas multas.

PATRÓN: Los datos genéricos (leyes, principios) suelen ser correctos. Las estadísticas específicas (%, cifras, estudios) son las más propensas a alucinación.

Tabla: fiabilidad de la IA por tipo de tarea

TIPO DE TAREA	FIABILIDAD	VERIFICACIÓN NECESARIA	EJEMPLO
Estructura y formato	Muy alta	Mínima	Organizar bullets, crear tabla
Redacción y estilo	Alta	Revisión de tono	Emails, posts, propuestas
Programación (lógica)	Alta	Ejecutar y testear	Funciones, algoritmos, scripts
Análisis y razonamiento	Media-alta	Verificar lógica	Pros/contras, comparativas
Datos específicos	Baja-media	Verificar SIEMPRE	Estadísticas, citas, fechas
Eventos recientes	Baja	Verificar SIEMPRE	Noticias, precios actuales
Medicina, legal, fiscal	Variable	Profesional obligatorio	Diagnósticos, leyes, impuestos

Prompts anti-alucinación: kit completo

PROMPT 1: Forzar honestidad

"Responde solo con información de la que estés seguro.

Para cada afirmación, clasifica como:

- ✓ VERIFICADO: dato bien establecido
- ~ PROBABLE: razonable pero no confirmable
- ? INCIERTO: podría ser inexacto

Si no tienes información fiable, responde 'No tengo datos fiables sobre esto' en lugar de inventar."

PROMPT 2: Separar fuentes

"Para cada dato estadístico o afirmación fáctica:

- Indica la fuente si la conoces
- Si no puedes citar una fuente verificable, marca como [FUENTE NO VERIFICABLE]
- No inventes referencias bibliográficas"

PROMPT 3: Pedir alternativas

"Dame 3 perspectivas diferentes sobre [tema].

Para cada una, indica: quién la defiende, qué evidencia la apoya y qué críticas recibe."

PROMPT 4: Verificación cruzada

"Acabo de generar este texto con otra IA: [pegar texto].

Revisa cada dato fáctico. Para cada uno, indica si:

- Lo puedes confirmar
- Te parece dudoso
- Es claramente incorrecto"

Workflow de verificación para publicación

ANTES DE PUBLICAR CUALQUIER CONTENIDO GENERADO POR IA:

- Paso 1: ¿Hay estadísticas específicas?
 - Buscar la fuente original en Google
 - Si no encuentras la fuente, eliminar el dato
- Paso 2: ¿Hay nombres de personas o instituciones?
 - Verificar que existen
 - Verificar que la afirmación atribuida es real
- Paso 3: ¿Hay URLs o enlaces?
 - Hacer clic en cada uno
 - Verificar que lleva donde dice
- Paso 4: ¿Hay código?
 - Ejecutarlo en un entorno de prueba
 - Verificar que funciona como se describe
- Paso 5: ¿Hay fechas o cronologías?
 - Verificar las fechas clave
 - Especialmente si son recientes (post-training)
- Paso 6: Lectura final de coherencia
 - ¿Algo "suena demasiado perfecto"?
 - ¿Hay afirmaciones sin matices donde debería haberlos?

Errores comunes al manejar alucinaciones

- **Confiar porque "suena bien":** Las alucinaciones son convincentes por diseño. La IA genera texto estadísticamente plausible, no necesariamente verdadero. Que suene profesional no significa que sea correcto.
- **Pedir fuentes y confiar en ellas:** La IA también puede inventar las fuentes. Un "según estudio de Harvard 2023" puede ser tan inventado como el dato que cita.
- **Asumir que RAG elimina alucinaciones:** RAG reduce alucinaciones al anclar respuestas en documentos reales, pero no las elimina. La IA puede interpretar mal el documento o mezclar información de varias fuentes.
- **No verificar el código generado:** El código también puede tener "alucinaciones": funciones que no existen, parámetros incorrectos o lógica que compila pero no hace lo esperado. Siempre ejecuta y testea.

Consejo: Desarrolla un "sexto sentido" para detectar alucinaciones. Las señales más fiables son: datos demasiado redondos (exactamente 73%), afirmaciones absolutas sin matices ("sin

duda alguna"), y exceso de detalle no solicitado (año, página y editorial de un libro que no pediste). Con el tiempo, aprendes a distinguir las respuestas sólidas de las fabricadas.

IAcademy · iacademy.com · © 2026 Riskitera S.L.U.

MT08: Alucinaciones · Avanzado

¿Dudas? Usa el tutor IA en iacademy.com/course/mt08/