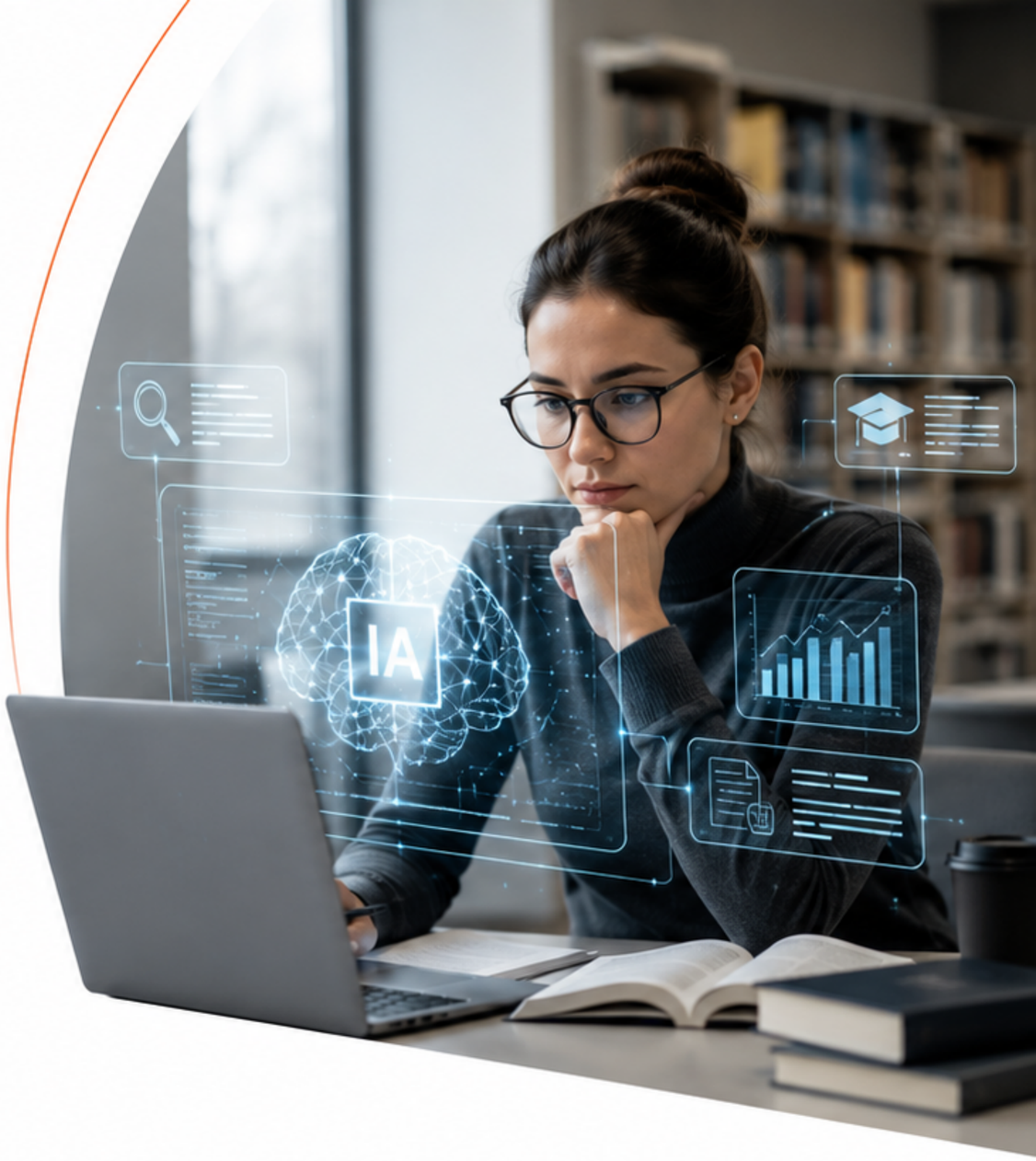
UDLA
UNIVERSIDAD DE LAS AMÉRICAS

Qué son las alucinaciones de la IA

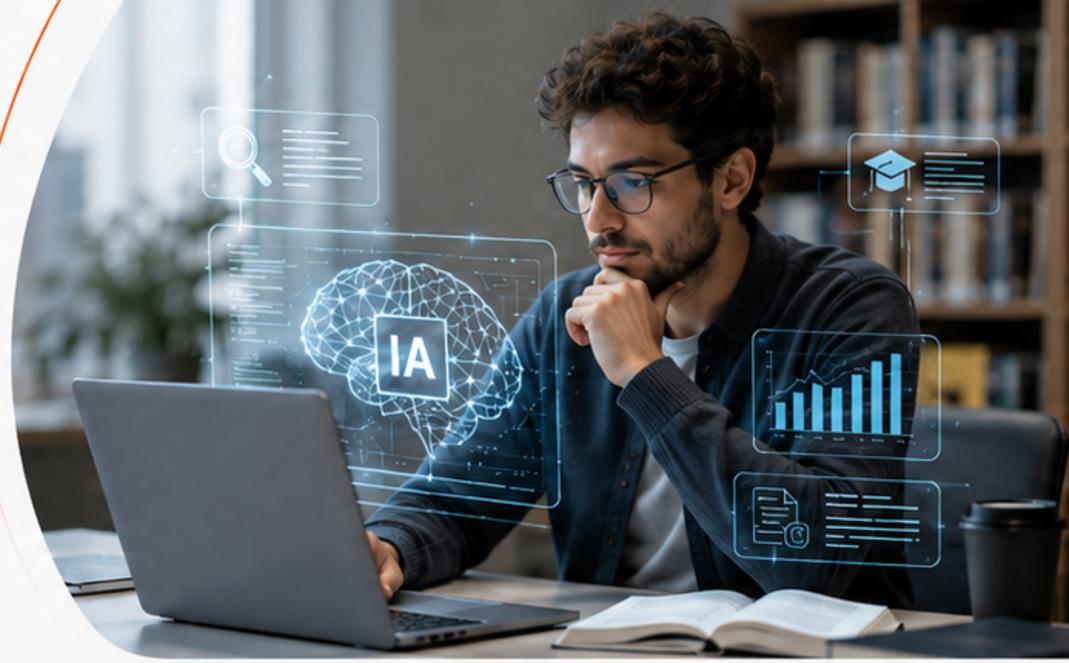
Definición, causas y estrategias para reducirlas



Ana Henríquez Orrego
Académica del Observatorio de IA en Educación

Por qué hablar de alucinaciones de la IA hoy

Un tema crítico en el uso responsable y efectivo de la IA en la educación y el trabajo profesional.



1. Uso extendido

La IA está presente en el estudio, la docencia, la investigación y el trabajo profesional.

2. Confianza aparente

Las respuestas generadas por IA son cada vez más fluidas y convincentes, aunque no siempre correctas.

3. Riesgo de error

Las alucinaciones pueden llevar a decisiones equivocadas, pérdida de tiempo y daños a la reputación.

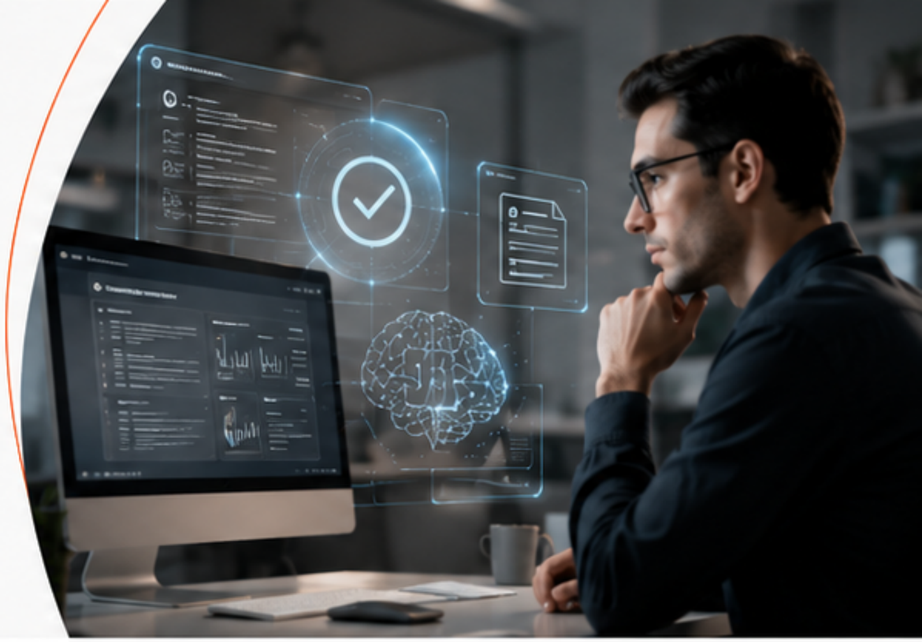
4. Necesidad de verificación

Verificar fuentes, contrastar información y desarrollar pensamiento crítico es más importante que nunca.

5. Relevancia para universidades

- Formar personas responsables y críticas en el uso de IA.
- Integrar alfabetización en IA en currículos y evaluaciones.
- Promover investigación y políticas institucionales claras.

Qué son las alucinaciones de la IA



Definición: Son respuestas que genera un sistema de inteligencia artificial y que, aunque parecen correctas y están bien formuladas, contienen información falsa, inventada, imprecisa o no verificable.

- Respuesta con apariencia de solidez y fluidez.
- Información falsa, inventada, imprecisa o no verificable.
- Puede incluir datos, citas o explicaciones incorrectas.
- No expresa intención de engaño, sino los límites del sistema.
- Más probable en temas complejos, nuevos o con poca evidencia.

Así se produce una alucinación

Respuesta
El modelo genera una respuesta fluida, coherente y bien estructurada.

+

Error factual o inferencial
La respuesta contiene información incorrecta, inventada o no sustentada en fuentes confiables.

+

Apariencia de veracidad
La redacción, el tono y los detalles hacen que la información parezca creíble y verificable.

=

Alucinación de IA
Se presenta como un hecho confiable, aunque sea falso, parcial o no verificable.

Cómo funciona este fenómeno

Las alucinaciones no son errores intencionales:
son el resultado de cómo los modelos generan texto.

Prompt

El usuario realiza una pregunta o solicita información.

Procesamiento

El modelo analiza patrones aprendidos en grandes volúmenes de datos.

Predicción probable

Genera la siguiente secuencia de palabras más grobable según esos patrones.

Respuesta plausible

El resultado puede sonar correcto y fluido, aunque no sea verídico o verificable.

Verificación humana

Es clave revisar, contrastar y validar la información antes de tomar decisiones o compartirla.

¿Por qué ocurren alucinaciones?

- La IA predice secuencias probables, no verdades absolutas.
- Aprende patrones estadísticos y relaciones lingüísticas, no hechos.
- Su fluidez verbal no implica comprensión del mundo real.
- Puede llenar vacíos con información plausible cuando faltan datos.
- Cada respuesta es probabilística y puede variar.

Por qué se producen las alucinaciones

Limitaciones del entrenamiento

Los modelos aprenden patrones, no la verdad absoluta.

Mayor riesgo en temas recientes o especializados

Hay menos datos disponibles y más cambios constantes.

Mezcla de patrones correctos con asociaciones erróneas

Combina información válida con inferencias incorrectas.

Cobertura desigual de datos

No todos los temas, regiones o perspectivas están bien representados.

Prompt ambiguo o poco preciso

Instrucciones poco claras aumentan la probabilidad de respuestas inexactas.

Falta de acceso a fuentes verificadas en algunos contextos

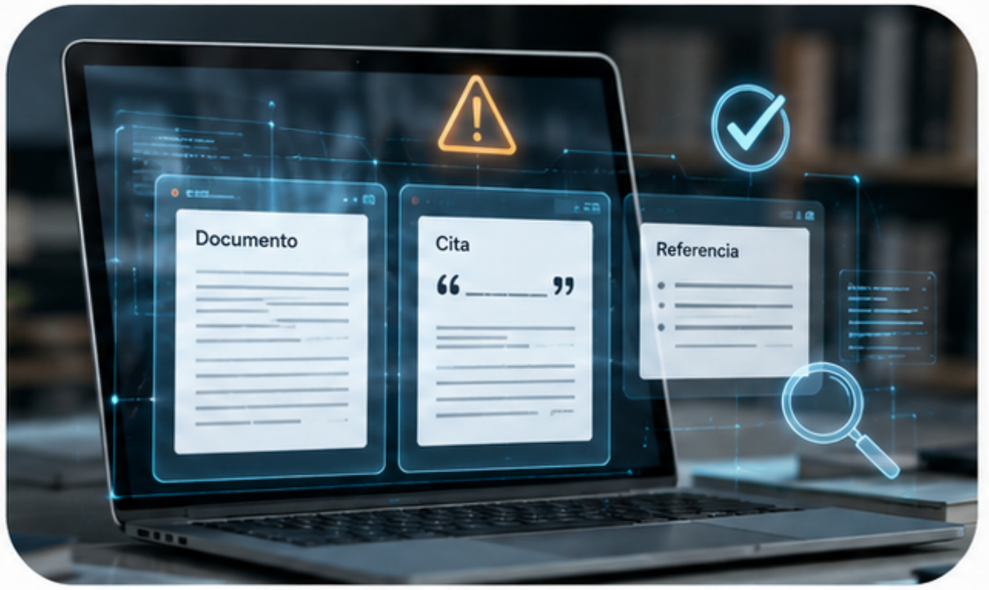
Sin verificación externa, la IA debe inferir.

Tendencia a responder aunque falte información

Busca dar una respuesta completa, incluso si debe rellenar vacíos.

Cómo se manifiestan las alucinaciones

Tipo frecuente	Ejemplo breve
Datos falsos	Indica que Chile tiene 54 millones de habitantes.
Citas o referencias inventadas	Atribuye una frase a un autor real, pero la cita no existe o está mal formulada.
Atribuciones incorrectas	Asigna una teoría o descubrimiento a una persona que no lo realizó.
Explicaciones conceptuales imprecisas	Simplifica en exceso un concepto complejo o mezcla ideas de teorías distintas.
Fusión de información	Combina datos de distintas fuentes o temas sin señalar la diferencia.
Exceso de seguridad expresiva	Responde con firmeza y tono categórico afirmando algo que es incorrecto o incierto.
Errores de contexto	Aplica información válida en otro contexto, tiempo, lugar o audiencia.



Cómo detectarlas

- **Verifica siempre:** contrasta datos, citas y cifras con fuentes confiables.
- **Pregunta por la fuente:** solicita referencias claras y comprobables.
- **Reformula y profundiza:** pide al modelo que aclare o justifique su respuesta.
- **Compara perspectivas:** revisa otras fuentes, autores o enfoques.
- **Observa el tono:** desconfía de respuestas demasiado categóricas sin evidencia.

Riesgos e implicancias en educación y trabajo académico

Ámbitos de impacto

DOCENCIA

- Riesgo de uso acrítico de IA en la planificación y materiales.
- Pérdida de tiempo en diseño pedagógico y retroalimentación.

APRENDIZAJE

- Dependencia que reduce esfuerzo y autonomía.
- Comprensión superficial y menor retención.

EVALUACIÓN

- Mayor dificultad para verificar originalidad y autoría.
- Retroalimentación menos significativa y personalizada.

INVESTIGACIÓN

- Generación de evidencia poco confiable o sesgada.
- Riesgos éticos y de integridad académica.

Implicancias generales

Riesgo para la calidad de trabajos e informes.

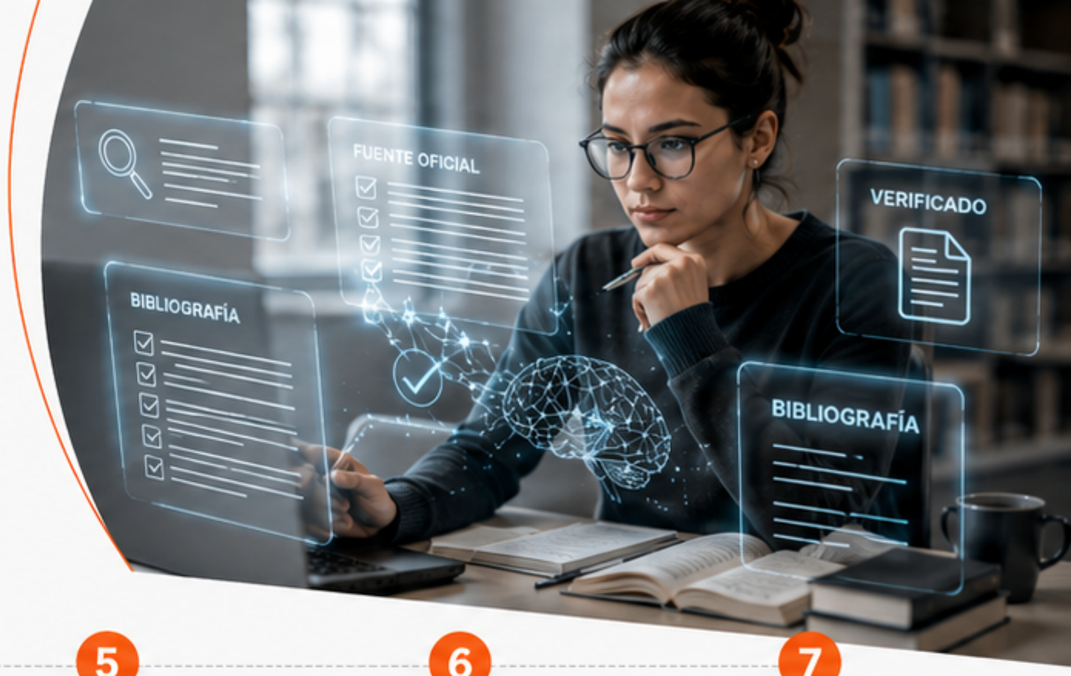
Impacto en la comprensión conceptual y el aprendizaje profundo.

Necesidad de pensamiento crítico y verificación sistemática.

Relevancia institucional para universidades: políticas claras, formación y cultura de uso responsable de la IA.

Cómo reducir las alucinaciones de la IA

Siete acciones prácticas para obtener respuestas más fiables, verificables y útiles.



1 Formula prompts claros y específicos

Define contexto, objetivo, audiencia, formato y límites para reducir interpretaciones.

2 Pide supuestos o nivel de certeza

Solicita que explicite supuestos, dudas y el grado de confianza de sus respuestas.

3 Solicita fuentes y verifiicalas

Pide referencias concretas y comprueba autoridad, fecha y relevancia.

4 Trabaja con corpus o documentos base

Entrega tus propios documentos y limita las respuestas a ese material.

5 Contrasta con fuentes oficiales o bibliografía

Compara con normas, reportes, artículos académicos o sitios oficiales.

6 Divide tareas complejas en pasos

Descompón el problema en preguntas o pasos secuenciales y valida cada resultado.

7 Mantén control humano en la validación final

Revisa, cuestiona y decide. Tu criterio es la última y más importante validación.

Ideas finales

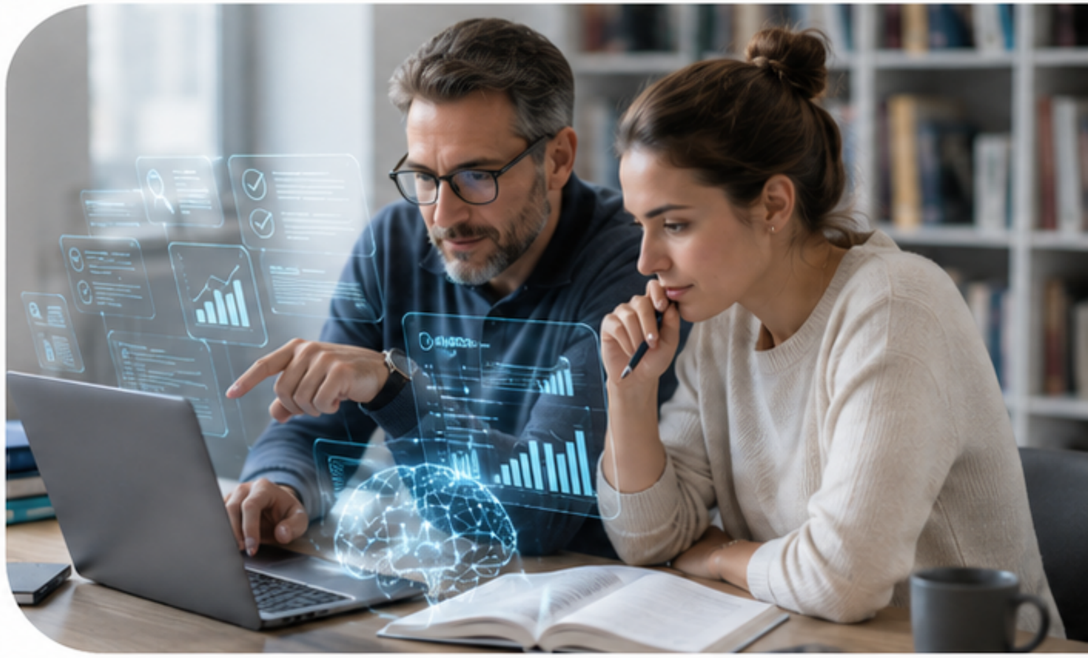
La IA es una gran asistente, no una fuente infalible.
Usarla con método crítico y buenas prácticas mejora la calidad y reduce riesgos.

Transparencia, trazabilidad y ética.
Cita, reconoce limitaciones y respeta la privacidad y la propiedad intelectual.

Protocolo práctico de verificación para docentes y estudiantes

Sigue estos pasos para evaluar y validar la información generada con IA.

1	**Leer críticamente**	Comprende el contexto, el propósito y los posibles sesgos de la respuesta.
2	**Identificar afirmaciones verificables**	Detecta datos, cifras, citas o relaciones que pueden comprobarse.
3	**Contrastar con fuentes confiables**	Verifica en al menos dos fuentes independientes, actuales y pertinentes.
4	**Revisar coherencia disciplinar**	Evalúa si la información se alinea con teorías, evidencia y estándares de tu área.
5	**Corregir, complementar o descartar**	Ajusta, amplía con evidencia o descarta lo que no sea válido.
6	**Declarar uso de IA cuando corresponda**	Indica cómo se usó la IA y cómo verificaste la información.
7	**Conservar el juicio humano como decisión final**	Tú eres responsable del criterio, la ética y el impacto de lo que compartes.



Preguntas clave para verificar

- ¿Qué afirma exactamente la respuesta?
- ¿Qué evidencia sustenta cada dato o idea?
- ¿Qué fuentes lo confirman o contradicen?
- ¿Es actual y pertinente para mi contexto?
- ¿Hay sesgos o vacíos importantes?
- ¿Cómo cambiaría mi conclusión con mejor evidencia?
- ¿Estoy listo para respaldar esta información públicamente?

Una cultura de uso crítico y responsable de la IA

- **Comprender las alucinaciones** fortalece una relación más madura con la IA.
- **La alfabetización en IA** integra conocimiento técnico básico, pensamiento crítico y verificación.
- **Las universidades** pueden formar usuarios con criterio y autonomía intelectual.
- **El valor pedagógico** crece con revisión, contraste y diálogo informado.
- Una cultura académica sólida requiere herramientas potentes y usuarios intelectualmente activos.

COMPRENSIÓN

Entiende cómo funciona la IA, sus límites y sus posibles errores.

VERIFICACIÓN

Contrasta, revisa y valida la información con fuentes confiables.

DECISIÓN HUMANA

Interpreta con criterio y decide con responsabilidad académica.

> **La IA puede apoyar el aprendizaje cuando el juicio humano guía la interpretación, la verificación y la decisión final.**