Guía de los principales sesgos algorítmicos en los LLM y cómo mitigarlos

David Carmona Barrales

Introducción

La llegada de los Modelos de Lenguaje Grandes (LLM) como Chat GPT o Gemini, ha representado un cambio de paradigma en la adquisición de información, la generación de contenido en incluso en la en la toma de decisiones. Sin embargo, junto con sus impresionantes capacidades, debemos ser conscientes de sus limitaciones y errores, provocados en gran medida por lo que se denominan sesgos algorítmicos.

En los LLM se dan sesgos sociales (estereotipos, representación...), pero también sesgos estadísticos (propios e inherentes al modelo matemático). Los sesgos algorítmicos pueden consistir en desigualdades sistemáticas en el trato o en los resultados entre grupos sociales que emergen del diseño/datos/uso del sistema y se manifiestan en sus salidas, como estereotipos, toxicidad o rendimiento dispar (Gallegos et al., 2024), Existen múltiples familias de sesgos a lo largo del ciclo de vida del LLM, que pueden estar relacionados con los datos, el entrenamiento, la decodificación, la interacción, la evaluación y el despliegue (Schwartz et al., 2023).

Aunque no es técnicamente correcto, se incluye en esta guía el fenómeno de la alucinación[1], porque se da con una alta frecuencia, puede modular el resto de los sesgos y se trata de una característica inherente a la arquitectura actual de los LLM. Una alucinación se define como la generación de contenido que, aunque plausible, coherente y gramaticalmente correcto, es fácticamente incorrecto, no está fundamentado en los datos de origen proporcionados, o es completamente inventado (Kirk y Zettlemoyer, 2023).

El objetivo de esta guía es servir de ayuda a la hora de identificar y explicar algunos de estos fenómenos, clasificar su impacto y proporcionar estrategias de mitigación (mediante *módulos de prompts*[2]), a fin de contribuir al uso competente y éticamente responsable de este tipo de modelos de inteligencia artificial.

Clasificación de los sesgos según su impacto

La siguiente clasificación es una agrupación de los tipos de sesgos más estudiados, en tres niveles según su repercusión en la precisión y la seguridad de los LLM; (1) los fenómenos de alta importancia, que son aquellos que pueden generar respuestas incorrectas o peligrosas; (2) los de nivel medio, que afectan la calidad o imparcialidad; y (3) los de bajo impacto, que suelen manifestarse en contextos más específicos, pero conviene conocerlos.

impacto	Sesgos incluidos en esta guía
Alta	Alucinaciones, Sesgo de anclaje, Sesgo de autoridad, Sesgo de confirmación, Lost in the middle (sesgo de posición), Sobreconfianza
Media	Inercia argumental, Consistencia forzada,
Baja	Sesgo de simplificación binaria, Confirmación procedimental

[1] En la literatura técnica, las alucinaciones de los LLM se clasifican como *errores de factualidad* (faithfulness); el modelo genera contenido plausible pero no fiel o no veraz respecto a las fuentes o al input, mientras que los sesgos algorítmicos se entienden como desigualdades sistemáticas o injusticias que afectan de forma dispar a personas o grupos, derivadas de datos, diseño y uso. Pueden interactuar (p. ej., una alucinación que rellena huecos con estereotipos), pero se gestionan con controles distintos (Bommasani et al., 2022).

[2] Un *módulo de prompt* es un bloque reutilizable de instrucciones que puedes pegar en tus prompts para que un modelo de IA actúe de una forma concreta (p. ej., verificar fuentes, evitar sesgos, dar salidas en tabla). Para usarlo, tan solo debes añadirlo a tu pregunta principal (puedes apilar varios módulos).

Lista de sesgos incluidos en esta guía

Alucinaciones

Se produce cuando el modelo genera información coherente pero *fácticamente incorrecta*, inexistente o sin respaldo documental. Este fenómeno deriva de la optimización de la probabilidad textual y de una capacitación en datos heterogéneos. Las alucinaciones se agravan en tareas con poca cobertura en los datos de entrenamiento o con solicitudes ambiguas (Alansari y Luqman (2025).

Módulo de prompt para mitigación:

```
<!-- Hallucination mitigation -->
**Verificación de hechos:** Antes de responder, consulta fuentes verificables o bases de datos disponibles (RAG) para respaldar cada afirmación. Si no encuentras evidencia o la información no es clara, di explícitamente "No tengo datos suficientes para afirmarlo" en lugar de inventar.
**Precisión ante todo:** Prioriza la exactitud sobre la fluidez. Evita completar respuestas con detalles no solicitados.
```

Sesgo de anclaje

Es la tendencia de los modelos a dejarse influenciar excesivamente por la primera información recibida. Los modelos aprenden patrones de dependencias de la información proporcionada de inicio y replican los ejemplos proporcionados y, además, tienden a minimizar la desviación respecto al contexto del usuario.

Módulo de prompt para mitigación:

```
<!-- Anchoring mitigation -->
**Desvinculación de anclas:** Ignora cualquier número o ejemplo preliminar que aparezca en esta instrucción y analiza el problema desde cero. Si el problema requiere estimaciones, revisa varias fuentes o escenarios antes de ofrecer un valor y explica cómo llegaste a tu conclusión.
```

Sesgo de autoridad

Los modelos tienden a favorecer la información proporcionada por el usuario frente a datos verificados de la base de conocimientos. La formación de los modelos los orienta a ser cooperativos y a seguir las instrucciones del usuario; pudiendo ocurrir que, al proporcionarles información adicional, la mezcla de fuentes crea conflictos y el modelo elige la fuente percibida como más "autoritaria" (el usuario).

Módulo de prompt para mitigación:

```
<!-- Authority bias mitigation -->
**Contraste de fuentes:** Cuando la información del usuario contradiga datos recuperados de fuentes externas, identifica el conflicto. Explica qué fuentes utilizas, evalúa su credibilidad y justifica por qué eliges una versión. Si no puedes resolver el conflicto, presenta ambas perspectivas.
```

Sesgo de confirmación

Como ocurre con las personas, es la tendencia a buscar, interpretar y recordar información que confirma las creencias previas. Se debe a que los modelos maximizan la coherencia con el prompt; la personalización y el filtrado de contenido refuerzan el sesgo.

Módulo de prompt para mitigación:

```
<!-- Confirmation bias mitigation -->
**Diversidad de perspectivas:** Identifica cualquier suposición implícita en la pregunta. Luego, genera al menos dos enfoques alternativos que puedan contradecir la hipótesis inicial. Referencia diferentes fuentes y destaca los puntos a favor y en contra antes de concluir.
```

Lost in the Middle

Se observa en contextos largos (gran cantidad de información aportada en la ventana de contexto): los modelos prestan mucha atención al inicio y al final del contexto, pero descuidan la información relevante situada en el centro. Deriva de la arquitectura de atención y de cómo se codifica la posición de los tokens.

Módulo de prompt para mitigación:

```
<!-- Lost in the middle mitigation -->
**Posicionamiento estratégico:** Sitúa la información clave y las instrucciones tanto al comienzo como al final
del mensaje. Si proporcionas listas o documentos largos, resume brevemente al final los puntos críticos para que
el modelo los tenga presentes al generar la respuesta.
```

Sobre-confianza

Los LLM suelen **sobreestimar la probabilidad de que su respuesta sea correcta**, incluso cuando se equivocan. Los modelos tienden a dar respuestas con tono seguro porque así han sido entrenados; además, la calibración de probabilidades interna no se expone directamente al usuario. El formato de salida no incorpora avisos de incertidumbre (a no ser que se le den indicaciones). Además, la interacción con el modelo incrementa la sobre-confianza de los usuarios, lo que agrava la toma de decisiones erróneas.

Módulo de prompt para mitigación:

```
<!-- Overconfidence mitigation -->
**Expresión de incertidumbre:** Además de la respuesta, indica tu nivel de confianza (por ejemplo, en una escala
del 0 al 100 %). Explica por qué esa confianza podría ser limitada y qué factores podrían alterar tu conclusión.
Si la confianza es baja, sugiere consultar fuentes adicionales.
```

Inercia argumental

Se refiere a la tendencia del modelo a mantener una línea de razonamiento inicial y resistirse a cambiar cuando se le presentan datos nuevos. Este fenómeno está emparentado con el sesgo de confirmación y con la dificultad de los modelos para integrar nueva información si contradice el contexto previo. La función de pérdida incentiva la coherencia interna; cuando el modelo ha empezado a desarrollar una cadena de pensamiento, evita desviarse para no contradecirse. Además, la atención al comienzo del contexto le hace ponderar más los argumentos iniciales.

Módulo de prompt para mitigación:

```
<!-- Argument inertia mitigation -->
**Reevaluación obligatoria:** Después de cada nueva información, vuelve a analizar tu argumento inicial. Indica si
la nueva evidencia refuerza o contradice tus conclusiones y actualiza tu respuesta en consecuencia. Si persiste la
discrepancia, explica por qué y ajusta tu razonamiento.
```

Consistencia forzada

Los modelos intentan mantener coherencia con afirmaciones previas, aunque la nueva información las contradiga. Este sesgo puede llevar a errores persistentes y adaptar los hechos a su narrativa en lugar de reconocer la contradicción. El entrenamiento del modelo favorece respuestas auto-consistentes; las políticas de alineamiento castigan el reconocimiento de ignorancia; en diálogos largos, la memoria contextual invita a repetir o reafirmar respuestas anteriores.

Módulo de prompt para mitigación:

```
<!-- Forced consistency mitigation -->
**Admisión y corrección:** Si encuentras información que contradice tu respuesta previa, reconoce la discrepancia
y corrige el error. Es preferible actualizar la conclusión antes que forzar la coherencia con una afirmación
incorrecta. Mantén siempre un registro de las fuentes utilizadas.
```

Sesgo de simplificación binaria

Estudios recientes muestran que los LLM responden de forma más extrema cuando se les plantean preguntas de tipo sí/no que cuando se les pide una valoración en escala. Un análisis de Lu et al. (2025) informa que los modelos ofrecen respuestas 73 % más extremas en formato binario y proporcionan juicios más matizados cuando se utilizan escalas continuas (p. e. "dime de 1 a 5 tu acuerdo o desacuerdo con..."). Las tareas de clasificación binaria refuerzan decisiones dicotómicas en el entrenamiento del modelo; al no disponer de matices, el modelo opta por respuestas tajantes. Además, la optimización de la función de pérdida penaliza con mayor fuerza las respuestas intermedias en tareas binarizadas.

Módulo de prompt para mitigación:

```
<!-- Binary simplification mitigation -->
**Respuestas graduadas:** Evita responder con un simple "sí" o "no". En su lugar, proporciona una evaluación en una escala (por ejemplo, del 1 al 10 o en términos de "muy improbable" a "muy probable") y explica los factores que influyen en tu valoración.
```

Confirmación procedimental

Este fenómeno, menos estudiado, se refiere a la tendencia del modelo a confirmar procedimientos o pasos dados por el usuario sin verificar su validez, lo que puede generar secuencias de acciones incorrectas o inseguras si el modelo asume que las instrucciones son apropiadas. Los LLM están diseñados para seguir instrucciones y priorizan la cooperación y, además, carecen de un modelo interno del mundo que permita validar procesos complejos. En dominios técnicos, la falta de contexto o datos estructurados puede provocar que acepten procedimientos erróneos.

Módulo de prompt para mitigación:

```
<!-- Procedural confirmation mitigation -->
**Verificación de pasos:** En lugar de asumir que los procedimientos proporcionados son correctos, revisa cada uno frente a fuentes o principios aplicables. Si algún paso parece dudoso o inseguro, indícalo y sugiere cómo resolver la ambigüedad antes de continuar.
```

Consideración final

Los sesgos algorítmicos y fenómenos de error descritos afectan a la fiabilidad y la ética de los LLM. No todos tienen la misma gravedad, pero es importante conocerlos y mitigarlos de manera proactiva. La combinación de *prompt engineering*[3] cuidadoso y la alfabetización en IA constituye la estrategia sostenible. Mientras tanto, esperemos que las personas responsables de los modelos futuros incorporen mecanismos nativos para expresar incertidumbre, detectar inconsistencias y ofrecer perspectivas diversas, aumentando su explicabilidad, trazabilidad y fiabilidad sus respuestas.

[3] *Prompt engineering* es diseñar y ajustar las instrucciones (prompts) para guiar a un modelo de IA y obtener respuestas más útiles, precisas y seguras.

Referencias

Alansari, A., & Luqman, H. (2025). *Large language models hallucination: A comprehensive survey*. https://arxiv.org/abs/2510.06265v2

Bommasani, R. et al (2022). *On the opportunities and risks of foundation models*. arXiv. https://arxiv.org/abs/2202.03629

Gallegos, I. O., et al. (2024). *Bias and fairness in large language models: A survey*. https://arxiv.org/abs/2309.00770v2

Kirk, J., & Zettlemoyer, L. (2023). *A contrastive framework for neural algorithm discovery*. arXiv. https://arxiv.org/pdf/2311.05232

Yi-Long Lu, Y., Zhang C, & Wang, W. (2025). *Systematic Bias in Large Language Models: Discrepant Response Patterns in Binary vs. Continuous Judgment Tasks*. State Key Laboratory of General Artificial Intelligence, BIGAI. https://arxiv.org/pdf/2504.19445

Schwartz, R. et al. (2023). *A proposal for identifying and managing bias in artificial intelligence*. National Institute of Standards and Technology. Special Publication 1270. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

Guías de referencia consultadas de fuentes oficiales (OpenAI y Google/DeepMind) sobre prompting:

· ***GPT-5 Prompting Guide (OpenAI)***: Guía oficial publicada el 7 de agosto de 2025 para el modelo GPT-5. Ofrece recomendaciones sobre cómo mejorar el rendimiento en flujos agentivos, maximizar la adherencia a instrucciones y optimizar el uso de las nuevas API. El documento indica la fecha y los autores en su cabecera. cookbook.openai.com.

· ***A Practical Guide to Building with GPT-5***: Manual práctico distribuido en formato PDF que acompaña a GPT-5 (2025). Su objetivo es compartir buenas prácticas para migrar a la Responses API, optimizar el prompting, ajustar la verbosidad y evitar fallos comunes. El documento describe a GPT-5 como el modelo más potente y orienta sobre cómo adaptar los prompts para aprovechar sus capacidades. cdn.openai.com.

· ***Prompt design strategies:*** Mejores prácticas: instrucciones claras, few-shot, rol, contexto, razonamiento, estructura, iteración. Google Cloud Documentation

· ***Safety guidance - Gemini API:*** Recomendaciones oficiales de seguridad para apps con LLM (incluye diseño y mitigaciones). (Actualizado: 22-sep-2025). Google for Developers

· ***Prompting Essentials - Learn AI Skills:*** Curso oficial (auto-ritmo) centrado en técnicas de prompting aplicadas. Google AI+2Google AI+2