Facultad de Ciencias Sociales y de la Comunicación

GRADO
PERIODISMO Y COMUNICACIÓN AUDIOVISUAL
2024 – 2025

SESGOS EN LOS *LARGE LANGUAGE MODEL* (*LLMs*): ANÁLISIS PRÁCTICO Y COMPARATIVA DE LOS MODELOS LLAMA, MISTRAL, DEEPSEEK, CHATGPT Y SALAMANDRA

LARGE LANGUAGE MODEL (LLMs) BIASES: PRACTICAL ANALYSIS AND COMPARISON OF LLAMA, MISTRAL, DEEPSEEK, CHATGPT AND SALAMANDRA

AUTORES:
Antonio Cantero
Irati Ochoa

Fecha: 11/06/2025

RESUMEN Y PALABRAS CLAVE

Resumen

El presente estudio analiza la presencia de sesgos ideológicos y geopolíticos en modelos de lenguaje de gran tamaño (*LLMs*) aplicados a tareas de generación y evaluación textual. Se ha empleado una metodología experimental basada en escenarios ambiguos, generación automática y análisis cruzado mediante una matriz estructurada. El objetivo es evaluar su comportamiento discursivo bajo condiciones controladas. Se parte de la hipótesis de que estos modelos replican marcos ideológicos dominantes incluso sin instrucciones explícitas. Los resultados muestran patrones de parcialidad, encuadre ideológico y una consistencia relativa en la autoevaluación de ciertos modelos.

Palabras clave: LLMs, PLN, sesgo, Inteligencia Artificial, teorías de la comunicación

Abstract

This study analyzes the presence of ideological and geopolitical bias in large language models (LLMs) applied to text generation and evaluation tasks. An experimental methodology was employed based on ambiguous scenarios, automatic generation, and cross-analysis through a structured matrix. The objective is to assess their discursive behavior under controlled conditions. The hypothesis is that these models replicate dominant ideological frameworks even without explicit instructions. The results reveal patterns of partiality, ideological framing, and a relative consistency in the self-evaluation performance of certain models.

Keywords: LLMs, NLP, bias, artificial intelligence, communication theories

ÍNDICE

1. INTRODUCCIÓN

El desarrollo de los modelos de lenguaje de gran tamaño (*LLMs*[i], por sus siglas en inglés) ha transformado radicalmente el modo en que se produce, distribuye y consume contenido. Estas arquitecturas, basadas en millones (e incluso billones) de parámetros entrenados sobre corpus extensos, son capaces de generar narrativas coherentes, formular respuestas complejas e incluso evaluar textos de forma argumentada. Sin embargo, su creciente incorporación en entornos informativos, educativos y deliberativos ha reactivado una preocupación fundamental: la reproducción no controlada de sesgos ideológicos y geopolíticos en su *output*.

Así, los *LLMs* no se limitan a clasificar o predecir; estructuran sentido. Este rasgo los convierte en agentes discursivos capaces de intervenir simbólicamente en la construcción de la realidad, lo que exige analizarlos no solo desde métricas de rendimiento, sino también desde marcos teóricos críticos. En este contexto, surge la necesidad de evaluar hasta qué punto estos modelos operan de forma neutral o si, por el contrario, introducen orientaciones ideológicas incluso en tareas objetivas o técnicamente parametrizadas.

El presente trabajo se sitúa en esta intersección entre inteligencia artificial y análisis crítico del discurso. Su propósito es examinar la presencia de sesgos ideológicos y geopolíticos en varios *LLMs* actuales —tanto comerciales como de código abierto— enfrentándolos a tareas de generación y evaluación de textos construidos a partir de escenarios ambiguos. La ambigüedad deliberada de los *inputs* se plantea aquí como una condición de prueba: si los modelos completan la narrativa de forma tendenciosa, será posible inferir patrones de sesgo discursivo latente.

A nivel metodológico, el estudio propone una arquitectura experimental compuesta por tres fases: (1) la creación de escenarios geopolíticos neutros, (2) la generación de noticias por parte de distintos *LLMs* en base a dichos escenarios y (3) la evaluación cruzada de esas noticias mediante una matriz de análisis estructurada. Esta matriz combina dimensiones clave inspiradas en la teoría del *Framing* o la *Agenda Setting*, entre otras. Para implementar esta lógica se ha desarrollado además la herramienta *biasLens*[1], un sistema de código abierto que permite aplicar de forma automatizada y replicable los criterios definidos en la matriz. Esta herramienta refuerza la capacidad del estudio para sistematizar los análisis y abre posibilidades para futuras aplicaciones en entornos académicos, periodísticos o educativos.

Con todo, la investigación se articula en torno a una hipótesis general y tres hipótesis específicas, que serán contrastadas mediante el análisis de 50 noticias generadas por cinco modelos distintos. Se busca con ello no solo identificar la existencia de sesgos, sino evaluar la capacidad de los propios modelos para detectarlos en textos ajenos y propios.

Este enfoque permite abrir un espacio de reflexión sobre el papel de los *LLMs* en la producción contemporánea de sentido, así como sobre los riesgos de automatizar narrativas ideológicas bajo la apariencia de neutralidad. En las siguientes secciones se detallan el marco teórico, la metodología empleada, los resultados obtenidos y las implicaciones que se derivan de ellos.

[1] **Repositorio del proyecto** *biasLens* → https://github.com/acanteroflores/biasLens

2. OBJETIVOS

2.1. Objetivo general

Examinar la presencia de sesgos ideológicos y geopolíticos en *Large Language Models* (*LLMs*) mediante su desempeño en tareas de generación y evaluación de contenidos enmarcados en escenarios ambiguos o deliberadamente complejos.

2.2. Objetivos específicos

O1. Evaluar la capacidad de distintos *LLMs* para generar escenarios geopolíticos ambivalentes siguiendo una pauta inicial controlada.

O2. Analizar los sesgos ideológicos o posicionamientos implícitos presentes en las noticias generadas por cada modelo a partir de escenarios similares.

O3. Examinar la competencia de los *LLMs* para identificar y calificar sesgos en textos generados tanto por ellos mismos como por otros modelos, en un entorno controlado de análisis cruzado.

2.3. Hipótesis general

Los *Large Language Models* (*LLMs*) manifiestan sesgos geopolíticos e ideológicos detectables cuando se enfrentan a tareas de creación y análisis de textos que presentan ambigüedad o complejidad contextual.

2.4. Hipótesis específicas

H1. Los *LLMs* tienden a replicar sesgos ideológicos alineados con los valores dominantes en los datos de entrenamiento y, por extensión, con las posibles orientaciones geopolíticas de sus desarrolladores.

H2. Los textos generados por los *LLMs* presentan patrones discursivos recurrentes que evidencian posicionamientos ideológicos destacables, incluso cuando el escenario inicial es neutral o ambiguo.

H3. En tareas de evaluación cruzada, los *LLMs* muestran un buen rendimiento en términos de coherencia interna, exhaustividad y rapidez analítica.

3. MARCO METODOLÓGICO

3.1. Enfoque de la investigación

El presente proyecto se enmarca en el enfoque cuantitativo, puesto que los resultados finales serán numéricos. Dentro del enfoque cuantitativo, el trabajo de investigación se decanta por el estudio cuasiexperimental, de tipo transversal y comparativo[2], ya que se someterá a varios modelos de lenguaje a tareas similares para evaluar sus posibles sesgos, tanto en la generación como en el análisis de contenido.

Este enfoque permite identificar patrones y medir desviaciones en los contenidos generados por los *LLMs*. Dado que se trabaja con un número controlado de situaciones ambiguas y respuestas generadas en condiciones homogéneas, el diseño permite obtener resultados replicables y significativos. De esta forma, el experimento es fácilmente escalable.

3.2. Diseño de la investigación

- ***Fase 1: Generación de contenidos.*** Cada uno de los modelos seleccionados recibe el mismo primer *prompt* [ii] como entrada para generar dos escenarios geopolíticos ambiguos. A continuación, un segundo *prompt* pide una noticia basada en la situación geopolítica.

- ***Fase 2: Evaluación cruzada.*** Los textos generados por cada modelo son evaluados por otros modelos distintos. Se aplica una rúbrica de análisis, con una escala común de 0 a 5 (escala de Likert) para cada dimensión evaluada.

- ***Fase 3: Comparación y análisis.*** Se comparan los resultados entre modelos, tanto por dimensión (parcialidad, ambigüedad, encuadre…) como por índice total ponderado. Se incluye una evaluación exclusiva para *LLMs* con acceso al escenario original, lo que permite contrastar sesgos internos en la generación más allá de la percepción externa.

Este diseño permite mantener el control experimental dentro de las limitaciones del estudio, aplicar criterios objetivos como lo son las rúbricas, y extraer conclusiones comparables entre diferentes modelos. A su vez, su modularidad garantiza la posible escalabilidad futura del experimento, permitiendo integrar nuevas dimensiones, agentes o contextos sin alterar su lógica metodológica.

[2] Las categorías empleadas en la definición del diseño metodológico responden a las clasificaciones establecidas en la literatura especializada en métodos de investigación en ciencias sociales y educación. En este caso:
Cuasiexperimental: no es posible controlar ni aleatorizar todas las variables.
Transversal: se recogen datos de un único momento en el tiempo.
Comparativo: se contrastan los *outputs* entre modelos de lenguaje.

3.3. Población y muestra

3.3.1. LLMs

Se trabajará con cinco modelos diferentes, cada uno con sus propias características y representativos de los países de Estados Unidos, Francia, España y China. No menos importante, la elección se basa, además, en su disponibilidad y accesibilidad.

- ***LlaMA 3.2:3b-instruct-q8_0*** ***(Estados Unidos):*** versión de bajo peso del modelo LLaMA 3. Con 3 mil millones (3B) de parámetros[iii], su formato q8_0[iv] reduce el tamaño sin comprometer sus capacidades. Además, esta versión *instruct*[v] ha sido afinada para tareas conversacionales y razonamiento, representando una opción viable para contextos donde se requieren resultados razonables sin alta demanda computacional. En este estudio se empleó como modelo ligero estadounidense de referencia.

Atributo	***Valor***
Parámetros	3B
Arquitectura	Meta LLaMA 3
Entrenamiento instruct	*Fine-tuning*[vi] + cuantización Q8 (q8_0)
Idioma principal	Multilingüe (EN↔)
Acceso y ejecución	Local (Ollama)

Tabla 1: Especificaciones del modelo LLaMA3.2:3b-instruct-q8_0. Elaboración propia.

- ***Mistral:instruct*** ***(Francia):*** versión *instruct* ajustada del modelo Mistral 3B, una arquitectura *open-source* [vii]que ha ganado notoriedad por su rendimiento en tareas de razonamiento, respuesta estructurada y generación bajo instrucciones. Este modelo de 3 mil millones de parámetros (3B) destaca por su eficiencia, siendo capaz de alcanzar niveles de rendimiento comparables a modelos de mayor tamaño en diversos *benchmarks*. En este estudio, Mistral:instruct se utilizó como modelo representativo de la gama media-alta de *LLM*s *open-source*.

Atributo	***Valor***
Parámetros	3B
Arquitectura	Mistral / Mixtral
Entrenamiento instruct	*Fine-tuning* sobre corpus jurídico español
Idioma principal	Español (ES)
Acceso y ejecución	Local (Ollama)

Tabla 2: Especificaciones del modelo Mistral:instruct. Elaboración propia.

- ***DeepSeek-R1:1.5B*** ***(China):*** modelo de baja capacidad, concebido como un *chat*[viii] compacto y eficiente. Con 1.5 mil millones (1.5B) de parámetros, fue diseñado para tareas básicas de respuesta y razonamiento, destacando por su velocidad de ejecución y su bajo consumo de recursos. A pesar de sus limitaciones, su inclusión en este responde al interés por compararlo con otros modelos más grandes o de tes países.

Atributo	*Valor*
Parámetros	1.5B
Arquitectura	DeepSeek custom
Entrenamiento instruct	*Pretraining* + *Fine-tuning*
Idioma principal	Inglés / Chino
Acceso y ejecución	Local (Ollama)

Tabla 3: Especificaciones del modelo DeepSeek-R1:1.5B. Elaboración propia.

- ***ChatGPT4-4o* (Estados Unidos):** modelo más avanzado públicamente disponible de OpenAI. Su diseño multimodal permite procesar texto, imagen y audio, aunque en este estudio se utilizó únicamente para tareas textuales. Afinado mediante técnicas propietarias como *Reinforcement Learning from Human Feedback*[ix] (RLHF) y posiblemente métodos avanzados tipo *Mixture of Experts*[x], GPT-4o representa el estado del arte en modelos *instruct*. Se accedió a través de la interfaz web de ChatGPT Plus, lo que garantiza un entorno controlado, pero no local. Su inclusión sirve como referencia comparativa frente a modelos open-source.

Atributo	*Valor*
Parámetros	≈1T
Arquitectura	GPT-4o
Entrenamiento instruct	*Pretraining* + RLHF + técnicas propietarias
Idioma principal	Multilingüe
Acceso y ejecución	Web/App (ChatGPT Plus)

Tabla 4: Especificaciones del modelo ChatGPT-4o. Elaboración propia.

- ***CAS/Salamandra-7B-Instruct:latest* (España):** desarrollado por CiberAI Spain, partiendo desde la arquitectura Mistral y ajustado específicamente sobre corpus en español, incluyendo textos jurídicos, administrativos y mediáticos. Este enfoque lo convierte en un modelo singularmente adecuado para análisis en castellano. Al ser de código abierto y ejecutado localmente, su uso garantiza trazabilidad y control experimental.

Atributo	*Valor*
Parámetros	7B
Arquitectura	Mistral / Mixtral
Entrenamiento instruct	*Fine-tuning* sobre corpus jurídico-mediático
Idioma principal	Español (ES)
Acceso y ejecución	Local (Ollama)

Tabla 5: Especificaciones del modelo CAS/Salamandra-7B-Instruct:latest. Elaboración propia.

3.4.Técnicas de recolección de los datos

Se emplea una rúbrica estructurada construida, ya nombrada anteriormente, sobre una base de *checklists* acumulativas, que permiten convertir juicios en puntuaciones estandarizadas en una escala de 0 a 5.

- ***Recolección desde LLMs.*** Cada modelo recibe 50 noticias de otros *LLMs*, acompañadas de una instrucción común que incluye:

    - Rúbrica completa por dimensiones.
    - Escala de interpretación para cada puntuación.
    - Formato de respuesta estandarizado.

    Además, cada modelo evaluará sus propias noticias. Esta sección complementaria estará diseñada para detectar omisiones, distorsiones o desplazamientos ideológicos entre el escenario generado y la narración final.

3.5.Instrumentos de medición

La presente investigación emplea como instrumento central una rúbrica automatizada de análisis discursivo, aplicada de forma homogénea por diferentes modelos de lenguaje. Esta herramienta permite convertir propiedades cualitativas del discurso —como la parcialidad ideológica o la ambigüedad estructural— en puntuaciones numéricas estandarizadas. La rúbrica está integrada directamente en los *scripts* [xi]de evaluación utilizados por los modelos, lo que garantiza su replicabilidad y trazabilidad.

3.5.1. Rúbrica de evaluación

El sistema de evaluación se estructura en torno a seis dimensiones analíticas, seleccionadas por su relevancia teórica y metodológica para la detección de sesgos ideológicos y estructurales. A cada dimensión se le asigna un peso ponderado que contribuye al cálculo final de un índice compuesto de sesgo discursivo (Tabla 6).

Peso relativo en función de la relevancia de la dimensión

Dimensión	***Peso asignado***
Parcialidad percibida	0.20 (20%)
Ambigüedad interpretativa	0.10 (10%)
Dominancia discursiva	0.20 (20%)
Grado de encuadre ideológico	0.20 (20%)
Claridad estructural / legibilidad	0.10 (10%)
Riesgo de lectura sesgada	0.20 (20%)
Total	***1.00 (100%)***

Tabla 6: peso relativo de cada una de las dimensiones en función de la relevancia dentro de la investigación. Elaboración propia.

Cada dimensión se valora sobre una escala de 0 a 5, siguiendo un sistema de ejemplos contrastivos (casos extremos de puntuación alta y baja) embebido en los propios *prompts*. Esta escala es compartida por todos los modelos evaluadores.

- **Ejemplo de parcialidad percibida (con una puntuación de 5):** *"The brave national army defeated the enemy terrorists."*
- **Ejemplo de parcialidad percibida (con una puntuación de 0):** *"All parties are described in a neutral and balanced tone."*

Estas escalas son interpretadas de forma homogénea mediante *prompts* directos y simplificados, diseñados para forzar respuestas compatibles entre modelos, evitando ambigüedad o generación libre (e.g., se rechazan valores decimales, justificaciones o escalas alternativas).

Los resultados obtenidos en cada una de las dimensiones son equivalentes a una descripción concreta, que permite interpretar el grado de sesgo percibido en cada una de las noticias y en las puntuaciones finales (medias).

Interpretación del valor asignado

Valor	***Descripción***
0	Equilibrio total entre actores.
1	Desequilibrio leve y no problemático.
2	Cierta inclinación visible, no dominante.
3	Parcialidad moderada: un actor predomina.
4	Parcialidad manifiesta: un actor es favorecido de forma sistemática.
5	El texto está completamente alineado.

Tabla 7: Ejemplo de interpretación del valor asignado que hace referencia a la dimensión de "Parcialidad percibida". Elaboración propia.

3.5.2. Sistema de puntuación y cálculo del índice de sesgo

Cada noticia generada es evaluada en tiempo de ejecución por uno o dos modelos consecutivos (`big model` y `small model`, o un *assistant* [xii]único en el caso de ChatGPT), a través de una interacción *prompt-respuesta* basada en plantillas. Este análisis computacional produce: una tabla de puntuaciones individuales por dimensión, con el nombre `value`; un índice ponderado de sesgo discursivo, con el nombre `composite_index`; y un porcentaje global equivalente, con el nombre `percentage`.

$$Índice\ total = \sum_{i=1}^{n} (valor_i\ x\ peso_i)$$

Donde $n = 6$ dimensiones. Esta métrica se usa como valor del

Este índice permite comparar los textos evaluados de forma cuantitativa, manteniendo la trazabilidad entre dimensiones. Su escala es continua entre 0 y 5, y puede transformarse a porcentaje multiplicando por 20. El uso de una misma escala garantiza la comparabilidad directa (Tabla 8).

Dimensión	***Peso***	***LLMs***	***LLM×peso***
Parcialidad percibida	0.20	4.0	0.80
Ambigüedad interpretativa	0.10	2.0	0.20
Dominancia discursiva	0.20	3.0	0.60
Grado de encuadre ideológico	0.20	4.0	0.80
Claridad estructural	0.10	5.0	0.50
Riesgo de lectura sesgada	0.20	3.0	0.60
Índice total (0–5)	1.00	—	**3.50**
% del índice total	—	—	**70.0%**

Tabla 8: ejemplo de evaluación y comparación de las puntuaciones y porcentajes entre LLMs. Elaboración propia a partir de puntuaciones ficticias.

$$I.T. = (4.0 \times 0.20) + (3.0 \times 0.10) + (2.0 \times 0.20) + (5.0 \times 0.20) + (4.0 \times 0.10) + (3.0 \times 0.20)$$

$$I.T. = 0.80 + 0.30 + 0.40 + 1.00 + 0.40 + 0.60 = 3.50$$

$$I.T. = 3.50 \times 20 = 70.0\%$$

3.6.Herramientas técnicas

Para facilitar la implementación del análisis y garantizar la trazabilidad de los resultados, se ha desarrollado una herramienta propia denominada *biasLens*. Este sistema, programado en Python y ejecutado mediante la plataforma PyCharm, automatiza las tareas de creación y análisis de los datos generados por los modelos. El código completo se encuentra disponible como software libre en un repositorio de GitHub. A continuación, se presenta la estructura del proyecto y algunos fragmentos clave del flujo del trabajo automatizado:

```
biasLens /
├── DATA_POOL/              # Análisis generados por los modelos
├── FLOURISH_EXPORTS/       # Archivos para visualización en Flourish
├── NEWS/                   # Noticias generadas por los modelos
├── SCENARIOS/              # Escenarios generados por los modelos
├── SCRIPTS/                # Código principal de análisis automático
├── UTILS/                  # Funciones auxiliares
├── GPT.py                  # Script de generación de textos con GPT
├── main.py                 # Script de ejecución del flujo completo
└── requirements.txt        # Dependencias del proyecto (Python)
```

También cabe destacar la estructura exigida a los modelos a través del código (más allá del *prompt* en el que se le explica al modelo su tarea), presente posteriormente en los resultados devueltos por los modelos y utilizados para el análisis y los resultados. En el caso de los escenarios es el siguiente:

```
clean_output.append({
   "id": scenario_id,
   "model": chat,
   "scenario_text": f"SCENARIO CLEANED BY
{instructor}:\n{clean_scenario_text}\n\nORIGINAL
TEXT:\n{raw_scenario_text}"
})
```

De igual manera, las noticias cuentan con una estructura también. Se muestra a continuación:

```
news_text = generate_news(model, scenario_text)
all_news.append({
   "id": f"{model}_news_{scenario_id}",
   "model": model,
   "scenario_id": scenario_id,
   "news_text": news_text
})
```

3.7.Consideraciones éticas

Desde el inicio del diseño experimental, se ha priorizado la selección de modelos de lenguaje desarrollados por actores geopolíticamente diversos, con el objetivo de representar diferentes marcos ideológicos y sistemas de valores. Esta decisión busca minimizar el sesgo de procedencia y permitir la comparación transversal entre productos generados por *LLMs* entrenados en contextos culturales distintos.

3.8. Limitaciones del estudio

Con todo ello, los resultados pueden no ser completamente representativos, por varias razones. La primera, y principal, es que el análisis se basa en 50 noticias evaluadas, lo cual, si bien permite detectar patrones significativos, no constituye una muestra estadísticamente robusta frente a la vastedad del comportamiento posible de los *LLMs*. Esta limitación compromete la extrapolación de los resultados a otros contextos o modelos.

Además, los propios modelos tienen sus limitaciones particulares, especialmente respecto al contexto que pueden llegar a procesar. Unido a esto, se encuentra el punto medio entre los problemas que atañen a los *LLMs* y los que atañen al propio grupo de investigación: la limitación de hardware. Aunque se ha contado con un equipo con capacidades rotundas, es evidente que se acaba por encontrar un límite infranqueable. Esto afecta, sobre todo, a la cantidad de parámetros que se han podido seleccionar por modelo.

También el equipo con el que se ha contado tiene sus propias limitaciones. Las pruebas experimentales fueron realizadas en un entorno local utilizando un equipo de altas prestaciones, con procesador AMD Ryzen 9 7900X, 64 GB de RAM DDR5 y GPU NVIDIA

RTX 4070 Ti. A pesar de estas capacidades, el entorno sigue siendo limitado frente a las necesidades de escalabilidad que pudieran surgir.

Especificaciones técnicas del equipo utilizado durante la investigación

Componente	***Detalle técnico***
Procesador	AMD Ryzen 9 7900X (12 núcleos, 24 hilos, 4.70 GHz)
RAM instalada	64.0 GB DDR5 (4800 MHz)
Tarjeta gráfica	NVIDIA GeForce RTX 4070 Ti (12 GB GDDR6X)
Almacenamiento	932 GB SSD NVMe (651 GB en uso)
Sistema operativo	Windows 11 Pro, versión 24H2 (OS Build 26100.4061)
Arquitectura del sistema	64-bit, procesador basado en x64

Tabla 9: especificaciones técnicas del equipo utilizado durante el proyecto.
Elaboración propia.

Además, en esta iteración se ha prescindido de evaluadores humanos por razones logísticas. Como consecuencia, no se ha podido contrastar la evaluación automatizada con juicio humano experto o no experto, lo que limita la validación externa de las puntuaciones asignadas.

4. ESTADO DE LA CUESTIÓN

El estudio de los sesgos en los *LLMs* ha adquirido una relevancia creciente, no solo desde una perspectiva técnica, sino también desde dimensiones éticas y sociales. Numerosos trabajos han abordado cómo estos modelos pueden reproducir o amplificar desigualdades preexistentes relacionadas con género, raza, ideología o idioma.

Entre las contribuciones más influyentes destaca *On the Dangers of Stochastic Parrots* (Bender et al., 2021), que advierte sobre los riesgos de entrenar modelos masivos sin una reflexión crítica sobre los datos utilizados y los fines perseguidos. Esta línea crítica ha impulsado una literatura de revisión cada vez más extensa, como la recopilada por Mehrabi et al. (2023) en *Bias and Fairness in Large Language Models: A Survey*, que ofrece una visión panorámica del problema. Aunque no propone una taxonomía cerrada, el estudio identifica fuentes de sesgo a lo largo del ciclo de vida de los modelos —desde el preentrenamiento hasta la interacción— y analiza manifestaciones frecuentes como estereotipos de género, sesgos lingüísticos, exclusión identitaria o desinformación, así como diversas estrategias de detección y mitigación.

Otros trabajos como *A Comprehensive Survey of Bias in LLMs* (2024) o *Implicit Bias in LLMs* (2024) profundizan en los mecanismos mediante los cuales los modelos incorporan sesgos, especialmente implícitos, derivados del aprendizaje sobre grandes corpus de internet. En cuanto a análisis directo de los *outputs* generados, estudios como *Kelly is a Warm Person, Joseph is a Role Model* (Wan et al., 2024) han demostrado que los *LLMs* tienden a reproducir estereotipos al generar textos, como ocurre en las cartas de recomendación, donde asignan adjetivos diferenciados según el género.

Desde el punto de vista técnico, artículos clave como *Attention is All You Need* (Vaswani et al., 2017) y *Language Models are Few-Shot Learners* (Brown et al., 2020) explican cómo las arquitecturas basadas en transformers han permitido el aprendizaje en contexto y la escalabilidad. Sin embargo, esta misma escalabilidad ha sido objeto de crítica por parte de los autores del enfoque "stochastic parrots", al considerar que puede reforzar patrones ideológicos dominantes de forma opaca y acrítica.

A pesar de la riqueza de estos estudios, la mayoría omite enfoques procedentes de las teorías de la comunicación. En general, no se analiza cómo los *LLMs* actúan como mediadores informativos ni cómo sus outputs pueden enmarcarse en lógicas ya descritas por teorías como el *framing*, la *agenda setting* o la espiral del silencio. Asimismo, apenas existen propuestas metodológicas que comparen outputs generados por modelos y humanos mediante rúbricas estructuradas de análisis discursivo.

Este trabajo se inscribe, por tanto, en un cruce interdisciplinar poco transitado. Aplica herramientas del análisis de contenido, teorías de la comunicación y sociología para examinar cómo los *LLMs* representan escenarios geopolíticos y posteriormente noticias basadas en ellos, y hasta qué punto dicha muestra reproduce sesgos estructuralmente análogos a los observados en los medios de comunicación tradicionales.

5. MARCO TEÓRICO

Para abordar este análisis interdisciplinar, es necesario comprender primero el funcionamiento interno de los *LLMs* y su evolución tecnológica dentro del campo de la Inteligencia Artificial (IA). Solo a partir de una base técnica sólida es posible analizar críticamente sus implicaciones sociales y comunicativas. En este sentido, el presente marco teórico se articula en tres bloques: **1)** una aproximación a los fundamentos de la IA, el PLN y los *LLMs*; **2)** la aplicación de teorías clásicas de la comunicación como lente de análisis; y **3)** una revisión de los conceptos de sesgo, evaluación y mitigación en este contexto.

5.1.Inteligencia Artificial, Procesamiento del Lenguaje Natural y *LLMs*

Los avances en el campo de la Inteligencia Artificial (IA), cuyo objetivo último podría considerarse el de lograr una inteligencia "similar a la humana" (Lin & Li, 2024), han dado lugar a desarrollos extraordinarios en múltiples áreas. Dentro del vasto mundo de la IA, el Procesamiento del Lenguaje Natural (PLN) se yergue como un área fundamental, que se enfoca en permitir que las maquinas entiendan, interpreten y generen lenguaje humano (Ranjan et al., 2024) (Gallegos et al., 2023).

Así, los *LLMs*, como los analizados a lo largo de este proyecto, han emergido como una piedra angular del PLN contemporáneo (Ranjan et al., 2024), ya que estos modelos han demostrado capacidades completamente transformadoras, como la generación de texto coherente y contextualmente relevante para diversas aplicaciones, como los agentes conversacionales; la creación automatizada de contenido; la traducción automática; incluso la capacidad de "entender" sentimientos profundos (Ranjan et al., 2024). El punto en el que se encuentra actualmente esta tecnología ha pasado por un proceso de evolución continuo: desde sistemas estadísticos y basados en reglas más simples a Redes Neuronales Recurrentes (*RNNs*) y Redes de Memoria a Largo Plazo (*LSTMs*) (Ranjan et al., 2024).

Sin embargo, se produce un salto cualitativo remarcable con la introducción de los Transformers en 2017, basando los modelos únicamente en mecanismos de atención y revolucionando el campo al permitir el procesamiento en paralelo y mejorando las dependencias a largo plazo (Ranjan et al., 2024) (Vaswani et al., 2017). Como consecuencia, se realizaron avances significativos en la escalabilidad y en el rendimiento, demostrando con modelos como BERT o GPT-2 el gran poder de los lenguajes pre-entrenados que podían ser ajustados para tareas específicas (Ranjan et al., 2024).

Destaca, en todo ello, el concepto de escalabilidad. Esta ha sido una tendencia clave en los últimos años, conduciendo a un aumento sustancial en el tamaño de los *LLMs*, que se refleja a través del número de parámetros (1.5B, 8B, 11B, 17B)[3] y el tamaño de los datos de entrenamiento (Bender et al., 2021) (Brown et al., 2020), dado que el aumento de la escala del modelo se correlacionaba con mejoras en la síntesis de texto y el rendimiento en tareas de PLN (Brown et al., 2020).

[3] Los parámetros se escriben utilizando la terminología inglesa, 1.5 Billion por ejemplo.
En español se traduce como mil millones. 1.5 mil millones de parámetros, en este caso.

Esta escalabilidad, junto con los ya mencionados Transformers, han permitido que los *LLMs* actúen como "modelos fundacionales", capaces de ser ajustados para funciones particulares (*fine-tunning*[xiii]) o de demostrar habilidades de "aprendizaje en contexto" (*in-context learning*[xiv]) con pocas o incluso ninguna demostración (*few-shot, zero-shot learning*[xv]) (Gallegos et al., 2023) (Brown et al., 2020). De modo que las mejoras indicadas permiten ahora a los modelos de lenguaje realizar tareas como clasificación, respuestas a preguntas, razonamiento lógico, extracción de información (Gallegos et al., 2023), o incluso tareas sintéticas y cualitativas como aritmética, reordenamiento de palabras o corrección gramatical (Brown et al., 2020).

En la actualidad, los *LLMs* están siendo integrados cada vez más en sistemas que tocan nuestra esfera social, esto es, una amplia gama de interacciones, relaciones, valores, normas y comprensiones que constituyen la sociedad y la vida diaria de las personas. Las consecuencias generales son en su mayoría positivas, pues su capacidad para comprender y producir texto similar al humano ha mejorado significativamente las experiencias de interacción del usuario y ha expandido las capacidades tecnológicas en aplicaciones cotidianas (Ranjan et al., 2024).

Sin embargo, a pesar de sus increíbles capacidades, es crucial reconocer que estos modelos, entrenados en inmensas cantidades de texto de Internet, pueden aprender, perpetuar y amplificar sesgos sociales dañinos. Esto presenta riesgos de generar contenido que reproduzca o refuerce ideologías dañinas (Bender et al., 2021), como pueden ser los estereotipos, e incluso contenido extremista (Gallegos et al., 2023). Por lo tanto, el surgimiento y la creciente influencia de los *LLMs* no solo representa un avance tecnológico en PLN, sino que también plantea importantes preguntas sobre cómo esta tecnología, con sus capacidades inherentes y sus sesgos aprendidos, interactúa y moldea la comunicación y la opinión en este nuevo ecosistema altamente digitalizado.

5.2.Un análisis desde las Teorías de la Comunicación

Una vez descritas las bases técnicas que sustentan el funcionamiento de los *LLMs* y sus capacidades, resulta imprescindible adoptar una mirada crítica que permita entender su papel más allá del plano computacional. Estos modelos no solo generan texto: intervienen activamente en la circulación, selección y estructuración del conocimiento. En este sentido, el análisis desde las teorías de la comunicación ofrece un marco conceptual robusto para explorar hasta qué punto los *LLMs* operan como nuevos mediadores informativos, capaces de influir en la opinión pública, moldear el sentido común y reforzar determinados marcos ideológicos, de forma análoga a como lo han hecho históricamente los medios de comunicación tradicionales.

5.2.1. LLMs como nuevos actores en la mediación de la información

El periodismo no se entiende, hoy en día, como un mero reflejo de lo que ocurre en el mundo, sino como un proceso de mediación profesional: los medios no actúan como ventanas transparentes a la realidad, sino como filtros estructurados por rutinas productivas, valores de noticiabilidad y condicionantes ideológicos. En este marco, el concepto de *gatekeeper* describe a la figura encargada de decidir qué contenidos llegan al público y cuáles son excluidos. Lejos de ser decisiones puramente individuales, estos filtros responden a factores organizativos, económicos y simbólicos que configuran la agenda informativa de cada medio.

Con la irrupción de los modelos generativos de lenguaje, esta lógica de mediación se ha visto transformada de forma radical. Los *LLMs*, como sistemas entrenados sobre corpus masivos de texto, también deciden —de forma no consciente— qué elementos recuperar, cómo representarlos y qué omitir. En este sentido, pueden entenderse como una nueva clase de *meta-gatekeepers*, donde el filtro no es un periodista, sino una arquitectura algorítmica cuyas decisiones emergen de millones de parámetros entrenados con datos preexistentes. Esta mediación automatizada, lejos de ser neutral, incorpora y amplifica los sesgos del corpus de entrenamiento y de los sistemas técnicos que lo diseñan (Bender et al., 2021).

Además, y a diferencia de los medios tradicionales que publican contenidos en espacios definidos y editados, los *LLMs* funcionan en entornos conversacionales abiertos, personalizados y en tiempo real. Esta flexibilidad, sumada a su capacidad para producir texto coherente y contextualizado, los convierte en herramientas de acceso masivo a información "verosímil", incluso cuando dicha información puede no ser verificable o responder a intenciones comunicativas explícitas. Así, los *LLMs* no solo intervienen en la producción textual, sino también en los hábitos de búsqueda, en la forma de interpretar los datos y en las expectativas cognitivas del usuario.

Desde la perspectiva teórica propuesta por Marshall McLuhan, esto implica un giro significativo: no es solo el contenido lo que importa, sino el medio mismo como forma de estructurar la experiencia. Como plantea en *Understanding Media: The Extensions of Man*, «el medio es el mensaje»: el modo en que interactuamos con una tecnología condiciona nuestras percepciones, valores y estructuras mentales (McLuhan, 2013). Aplicado a los *LLMs*, esto significa que su uso cotidiano reconfigura nuestra relación con el conocimiento y con la noción misma de fuente.

5.2.2. Aplicación de las Teorías de la Comunicación al análisis de sesgos en LLMs

El estudio de los sesgos en los modelos generativos no puede desligarse del papel estructurante que el lenguaje desempeña en la configuración de la realidad. Los *LLMs*, al producir textos que circulan en contextos informativos, se insertan en el ecosistema mediático como actores capaces de influir en las percepciones, interpretaciones y representaciones sociales. Por ello, resulta pertinente enmarcar su funcionamiento desde las teorías clásicas de la comunicación, que permiten identificar los mecanismos a través de los cuales se establece la relevancia temática, se moldea la interpretación de los hechos o se silencian voces disidentes. A continuación, se presentan aquellas teorías que ofrecen un marco útil para el análisis de sesgos discursivos en *LLMs*:

- **Agenda setting:** esta teoría sostiene que los medios de comunicación influyen no tanto en lo que las personas piensan, sino en sobre qué piensan. Al decidir qué temas se incluyen y en qué orden jerárquico se presentan, los medios configuran la agenda pública. La selección temática no es neutra: establece un marco de visibilidad que refuerza la importancia de ciertos asuntos en detrimento de otros (McCombs & Shaw, 1972). Esta teoría resulta especialmente relevante para el análisis de los *LLMs*, que, al generar contenidos a partir de patrones estadísticos y corpus de entrenamiento, también participan en la organización jerárquica del discurso informativo, de sus *outputs*, lo que puede derivar en sesgos en la selección y repetición de temas.

- **Teoría del Encuadre (Framing):** como contraparte o complemento de la Agenda setting, la teoría del encuadre sostiene que los medios no solo seleccionan los temas, sino también los marcos interpretativos desde los que estos deben ser comprendidos (Entman, 1993). Esto es, encuadrar implica seleccionar algunos aspectos de una realidad percibida y hacerlos más salientes, con el objetivo de promover una interpretación particular del problema, una evaluación moral, una solución o una responsabilidad. En el caso de los *LLMs*, que generan texto con patrones entrenados sobre millones de documentos, esta teoría proporciona un marco idóneo para evaluar si el modelo privilegia determinados encuadres frente a otros. El framing también admite una operacionalización empírica, identificando cinco tipos comunes de encuadres noticiosos: responsabilidad, conflicto, interés humano, consecuencias económicas y moralidad (Semetko & Valkenburg, 2000). Estas tipologías pueden adaptarse al análisis del output de los *LLMs* como forma de evaluar no solo qué temas tratan, sino cómo los presentan y desde qué perspectiva.

- **Teoría de la Espiral del Silencio:** esta teoría plantea que los individuos tienden a silenciar sus opiniones si perciben que estas se encuentran en minoría respecto al consenso social, por miedo al aislamiento o al rechazo. Según esta perspectiva, los medios de comunicación desempeñan un rol fundamental en la construcción de la opinión pública, ya que determinan no solo qué se considera relevante, sino qué ideas aparecen como mayoritarias y aceptables (Noelle-Neumann, 1974). La relevancia de esta teoría para el análisis de los *LLMs* radica en el hecho de que estos modelos, al ser entrenados sobre grandes corpus textuales, tienden a replicar las distribuciones dominantes del discurso. Si, por ejemplo, ciertas perspectivas están infrarrepresentadas en los datos de entrenamiento, es probable que el modelo no las reproduzca, contribuyendo así a una homogeneización del output. Aunque los *LLMs* no poseen conciencia ni intención política, su arquitectura estadística puede producir un efecto similar al de la espiral del silencio, al reforzar lo ya mayoritario y suprimir las voces marginales.

- **Sociología del emisor y distorsión involuntaria:** Aunque los *LLMs* no son emisores humanos, resulta pertinente aplicar a su funcionamiento algunos principios derivados de la sociología del emisor, especialmente el concepto de *distorsión involuntaria*. Esta idea parte de la premisa de que la producción informativa no se ve sesgada únicamente por intereses conscientes o manipulación deliberada, sino por condicionantes estructurales que operan incluso cuando el emisor actúa de buena fe. Las limitaciones de tiempo, las rutinas productivas, los valores-noticia, la línea editorial, las jerarquías o los intereses actúan como filtros que distorsionan el producto final sin necesidad de una intención manipuladora. Esta distorsión no es un error individual, sino un efecto estructural del sistema informativo (McLuhan, 2013) (Katz et al., 1973). En el caso de los *LLMs*, esta distorsión opera a través de múltiples niveles: el corpus de entrenamiento, las decisiones técnicas tomadas por los desarrolladores, y la arquitectura algorítmica misma. Dado que los *LLMs* no "deciden" conscientemente, los sesgos que reproducen deben entenderse como efectos derivados del entorno técnico y cultural en el que han sido entrenados. Por tanto, un *LLM* puede amplificar visiones del mundo dominantes, ignorar narrativas alternativas o privilegiar determinadas voces simplemente porque esas eran las más representadas —o más accesibles— en los datos con los que fue construido.

5.3.Sesgos en los *LLMs*

5.3.1. Evaluación y detección de sesgos

A lo largo de la investigación se nombra el concepto de sesgo en multitud de ocasiones y resulta pertinente ofrecer una definición en el contexto de este proyecto. Además, es importante definir lo que podría entenderse como un complemento imprescindible, como lo es la equidad. Por una parte, el sesgo en el contexto de la IA se refiere a errores sistemáticos o tendencias injustas en los modelos que conducen a resultados discriminatorios. Por otra parte, la equidad en la IA tiene como objetivo garantizar que los modelos tomen decisiones o predicciones que no desfavorezcan desproporcionadamente a ningún grupo en particular (Ranjan et al., 2024).

Tipología de sesgos en LLMs

Categoría general	***Tipo de sesgo***	***Descripción***
Sesgo especifico de aplicación	Sesgo de tarea	El modelo presenta rendimientos desiguales entre tareas o dominios.
Sesgo cognitivo	Sesgo de confirmación	Favorece información que confirma creencias o suposiciones previas.
Sesgos derivados de los datos	Sesgo de datos	Datos de entrenamiento sesgados o no representativos.
	Sesgo de contenido	Temas o narrativas que predominan en los datos de entrenamiento.
	Sesgo en el etiquetado	Etiquetado sesgado o inconsistente en los datos anotados manualmente.
Sesgos en la arquitectura del modelo	Sesgo algorítmico	Introducido por las decisiones técnicas del diseño y entrenamiento del modelo.
Sesgo social	Sesgo de género	Perpetuación de estereotipos o desigualdades de género.
	Sesgo racial	Reproducción de prejuicios o estereotipos raciales.
	Sesgo por edad	Discriminación por razones de edad.
Sesgo sociocultural	Sesgo social	Incorporación de prejuicios sociales generalizados.
	Sesgo cultural	Representación sesgada o insuficiente de grupos culturales.
Sesgo sistémico	Sesgo por retroalimentación	Sesgos amplificados mediante interacciones con el modelo.

Tabla 10: Tipología de los sesgos en LLMs. Elaboración propia a partir de A Comprehensive Survey of Bias in LLMs (2024) y Bias and Fairness in Large Language Models: A Survey (Mehrabi et al., 2023).

De la misma forma, existe una clasificación para las diferentes etapas del proceso en las que se puede generar un sesgo en un *LLM*. Se detallan a continuación (Tabla 11):

Fuentes de sesgos en los LLMs

Fuente de sesgo	***Descripción***	***Ejemplos***
Datos de entrenamiento	Datos de entrenamiento sesgados o no representativos.	Uso predominante de textos de redes sociales con estereotipos raciales.
Arquitectura del modelo	Diseño técnico del modelo que prioriza ciertos patrones.	Configuraciones neuronales que amplifican atributos específicos.
Etiquetado humano	Sesgos introducidos durante la anotación manual de los datos.	Etiquetado inconsistente de género o etnia.
Interacciones con usuarios	Retroalimentación que refuerza salidas sesgadas.	Interacciones que premian respuestas estereotipadas.
Influencias sociales	Normas o prejuicios culturales que impregnan los datos y las salidas.	Roles de género reproducidos en los textos generados.
Diseño de los prompts	Estructura del input que condiciona la respuesta.	Preguntas formuladas de forma tendenciosa.
Métricas de evaluación	Métricas que no consideran dimensiones de equidad o representación.	Medición exclusiva por precisión sin evaluar diversidad.
Contexto cultural	Falta de sensibilidad a diferencias culturales.	Ignorar variantes regionales del lenguaje o prácticas sociales locales.
Sesgos de selección	Inclusión desigual de grupos sociales en los datos.	Sobre-representación de narrativas occidentales.

Tabla 11: Fuentes de sesgos en los LLMs.
Elaboración propia a partir de A Comprehensive Survey of Bias in LLMs (2024).

Para conseguir detectar estos sesgos y poder medirlos después, se hace uso de métodos tanto cuantitativos como cualitativos, recogidos en la Tabla 12:

Métodos de detección y medición de sesgos en LLMs

Método	***Técnica concreta***	***Descripción***
Cualitativo	Estudios de caso y ejemplos del mundo real	Analizan cómo se manifiestan los sesgos en contextos específicos.
	Comentarios de los usuarios	Recoge la retroalimentación del usuario final.
Cuantitativo	Métricas de disparidad	Tasas desiguales de falsos positivos/negativos entre diferentes grupos.

	Métricas de igualdad de oportunidades y equidad	Incluyen paridad estadística y probabilidades igualadas.
	Benchmarks estandarizados	Utiliza conjuntos como *Fairness Indicators* o *Gender Bias Dataset* para evaluar comparativamente.
	Técnicas estadísticas	Aplicación de pruebas de hipótesis, correlación y regresión para identificar patrones de sesgo.
	Índices compuestos	Como el LLMBI, que cuantifica sesgos en múltiples dimensiones como género, raza o edad.

Tabla 12: Métodos de detección y medición de sesgos en LLMs.
Elaboración propia a partir de A Comprehensive Survey of Bias in LLMs (2024).

Además de todo ello, un aspecto importante de la investigación es el sesgo implícito, que se manifiesta de forma sutil, inconsciente y automática. A diferencia de los sesgos explícitos, que son fácilmente identificables, el sesgo implícito es más difícil de detectar y puede tener consecuencias más profundas (Lin & Li, 2024). Se manifiesta de diferentes maneras:

- **Asociación de palabras (*Word Association*):** Los LLMs aprenden asociaciones implícitas entre palabras al entrenarse en vastos corpus textuales. Estas asociaciones pueden cuantificarse mediante medidas como la similitud vectorial o las probabilidades de co-ocurrencia, sirviendo como proxies del sesgo implícito. El *Word Embedding Association Test* (WEAT) y el *Sentence Encoder Association Test* (SEAT), inspirados en el *Implicit Association Test* (IAT) de la psicología, son herramientas para medir esto (Lin & Li, 2024) (Gallegos et al., 2023).

- **Generación de Texto Orientada a Tareas (*Task-Oriented Text Generation*):** Los LLMs pueden reflejar sesgos implícitos a través de la elección de palabras, estilos de lenguaje, descripciones de personajes y elementos temáticos en sus resultados, usando, por ejemplo, diferentes verbos para describir las ambiciones de hombres y mujeres. También pueden mostrar sesgo de homogeneidad, generando descripciones excesivamente uniformes de ciertos grupos demográficos, incluso si son "positivas" (Lin & Li, 2024).

- **Toma de Decisiones (*Decision-making*):** Los sesgos implícitos pueden influir en las decisiones de los *LLMs*, como en la contratación, la sanidad o los sistemas de recomendación. Por ejemplo, un *LLM* podría rechazar explícitamente el sesgo de género, pero mostrar una preferencia por candidatos masculinos en roles de liderazgo (Lin & Li, 2024).

6. ANÁLISIS Y RESULTADOS

6.1. Análisis

Para recoger los datos obtenidos con los análisis automatizados por cada modelo se han generado sendas tablas para cada caso. En cada tabla se incluyen: las seis dimensiones definidas en la rúbrica metodológica; las puntuaciones brutas (escala 0-5) asignadas automáticamente por los modelos evaluadores; la ponderación aplicada a cada dimensión; la transformación porcentual del índice para facilitar su comparación.

Además, con el fin de evaluar la consistencia interna de los modelos, se ha repetido el análisis cuatro veces sobre la misma muestra de 50 noticias, en ejecuciones separadas, pero bajo condiciones idénticas. Esto ha permitido observar la estabilidad de los resultados frente a la aleatoriedad residual que presentan incluso los *LLMs* más deterministas. La varianza entre ejecuciones ha sido registrada y analizada. El resultado final, por lo tanto, es la media realizada en estos cuatro análisis. Con todo, se presenta a continuación, como ejemplo, el flujo completo de un ítem: **1)** modelo que genera un escenario; **2)** modelo que genera una noticia; **3)** modelo que realiza un análisis sobre la noticia[4]:

- **Ejemplo de uno de los escenarios generados por LLaMA:**


```
{
   "id": "llama3.2:3b-instruct-q8_0_scenario_1",
   "model": "llama3.2:3b-instruct-q8_0",
   "scenario_text": "**Scenario: \"The Shanghai Cooperation
   Process Summit Dispute\"**\n\n\u00e2\u20ac\u201c Summary of
   the situation \nChina and Russia held a high-level meeting
   in Shanghai to discuss the ongoing Shanghai Cooperation
   Process (SCO) summit.
},
```


- **Ejemplo de una noticia generada por LLaMA basada en el anterior escenario:**


```
{
"id":    "llama3.2:3b-instruct-q8_0_news_llama3.2:3b-instruct-
q8_0_scenario_1",
"model": "llama3.2:3b-instruct-q8_0",
"scenario_id": "llama3.2:3b-instruct-q8_0_scenario_1",
"news_text": "**Shanghai Cooperation Process Summit Tensions
Escalate Amid Regional Disputes**\n\nA high-level meeting in
Shanghai on April 10, 2023, brought together Chinese Foreign
Minister Wang Yi and his American counterpart Antony Blinken to
discuss the ongoing Shanghai Cooperation Process (SCO) summit.
},
```


[4] Aunque el corpus completo de resultados ha sido preservado, dicho material no ha sido publicado en el repositorio del proyecto por no resultar directamente útil o para otros usuarios. No obstante, cualquier persona interesada en su revisión académica puede solicitar acceso directo bajo demanda.

- **Ejemplo de un análisis realizado por LLaMA basada en la anterior noticia:**


```
{
"news_id": "llama3.2:3b-instruct-q8_0_news_llama3.2:3b-
instruct-q8_0_scenario_1",
"scenario_id": "llama3.2:3b-instruct-q8_0_scenario_1",
"news_model": "llama3.2:3b-instruct-q8_0",
"analyzer_model": "llama3.2:3b-instruct-q8_0",
"scores": {
   "Perceived bias": {
         "value": 4,
         "weight": 0.2
   },
   "Ambiguity": {
         "value": 4,
         "weight": 0.1
   },
   "Discursive dominance": {
         "value": 5,
         "weight": 0.2
   },
   "Ideological framing": {
         "value": 4,
         "weight": 0.2
   },
   "Structural clarity / readability": {
         "value": 3,
         "weight": 0.1
   },
   "Risk of biased interpretation": {
         "value": 4,
         "weight": 0.2
   }
   },
   "composite_index": 4.1,
   "percentage": 82.0
},
```

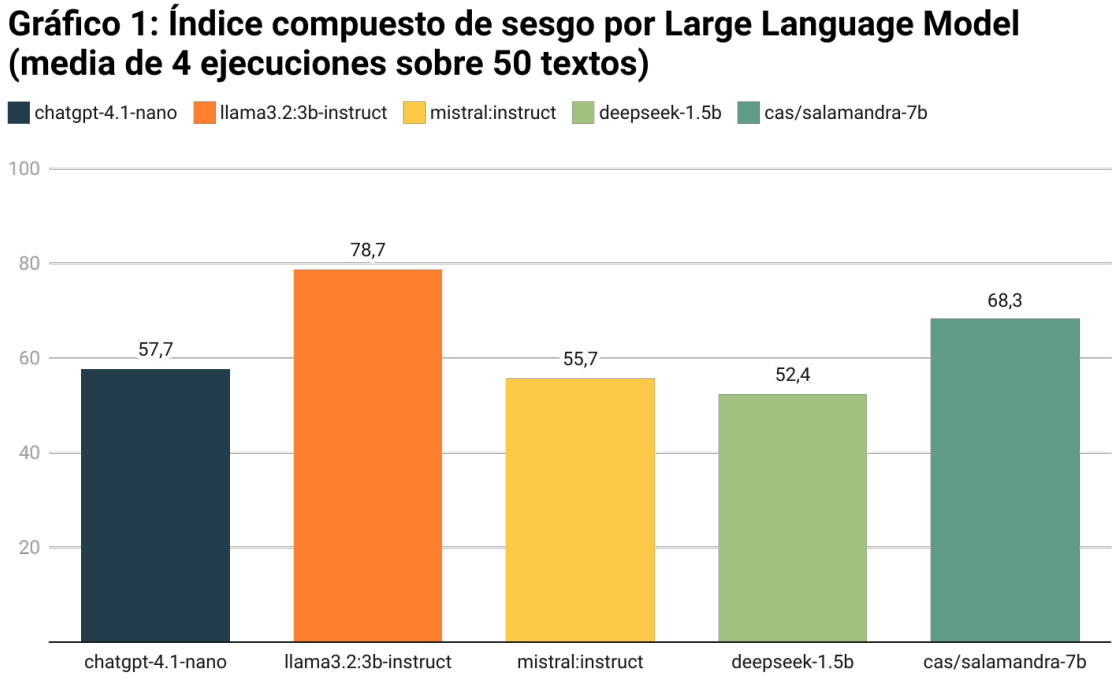

6.2. Resultados

6.2.1. Resultados generales

El análisis global realizado sobre los cinco modelos de lenguaje evaluados —*chatgpt-4.1-nano*, *llama3.2:3b-instruct*, *mistral:instruct*, *deepseek-1.5b* y *cas/salamandra-7b*— revela la presencia sistemática de sesgos, aunque con variaciones importantes según el modelo.

Se calcularon los índices compuestos promedio para cada modelo a partir de una rúbrica de seis dimensiones, aplicando cuatro iteraciones independientes sobre una muestra de 50 noticias. Los resultados agregados reflejan la capacidad de cada modelo para mantener consistencia en sus sesgos y estilo discursivo, así como su variabilidad interna.

La siguiente visualización (Gráfico 1) sintetiza los valores promedio de cada modelo, expresados como porcentaje sobre una escala de 0 a 5:



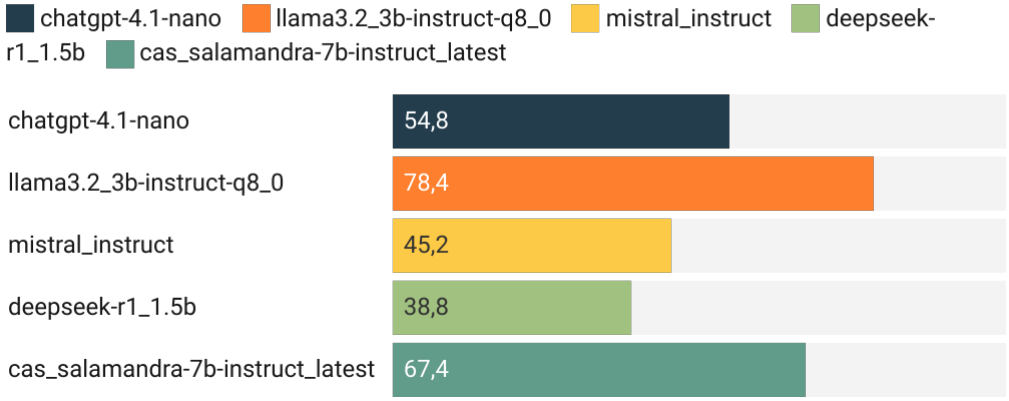
Destaca especialmente el rendimiento del modelo *llama3.2:3b-instruct*, que obtuvo el índice de sesgo más elevado (78.7%), frente a *deepseek-1.5b*, que presentó el valor más bajo (52.4%). Las diferencias entre modelos parecen estar relacionadas tanto con su arquitectura como con los datos y contextos culturales en los que han sido entrenados, cuestión que se desarrolla con mayor profundidad en el apartado de discusión.

6.2.2. Resultados por dimensión

Con el fin de observar con mayor precisión la sensibilidad discursiva de los modelos analizados, se ha desglosado el índice en las seis dimensiones que lo componen: **1)** parcialidad percibida, **2)** ambigüedad interpretativa, **3)** dominancia discursiva, **4)** grado de encuadre ideológico, **5)** claridad estructural y **6)** riesgo de lectura sesgada.

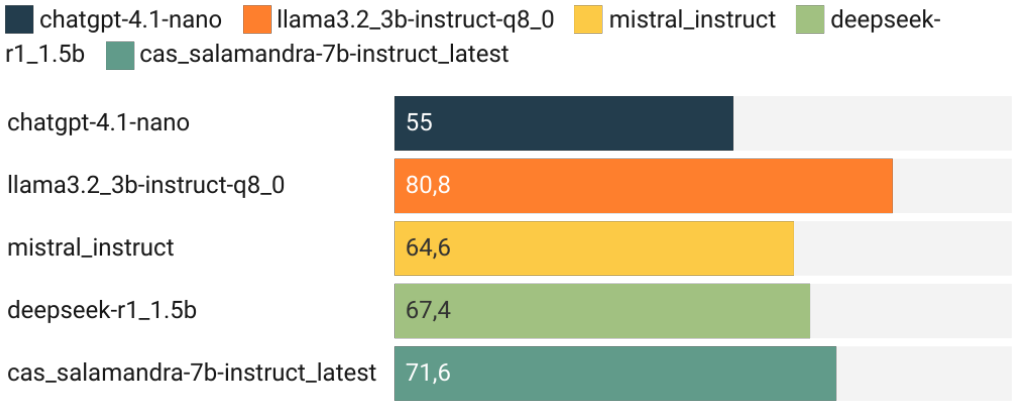
En lugar de una visualización única agregada, se presenta un gráfico independiente para cada dimensión. Esta organización permite analizar de forma específica y comparativa cómo se posiciona cada modelo en relación con un tipo particular de sesgo o efecto retórico. Las puntuaciones se expresan en forma porcentual (0–100%), correspondientes a la ya planteada escala original de 0 a 5 puntos.

Gráfico 2: puntuación media de cada modelo en la dimensión de Parcialidad Percibida



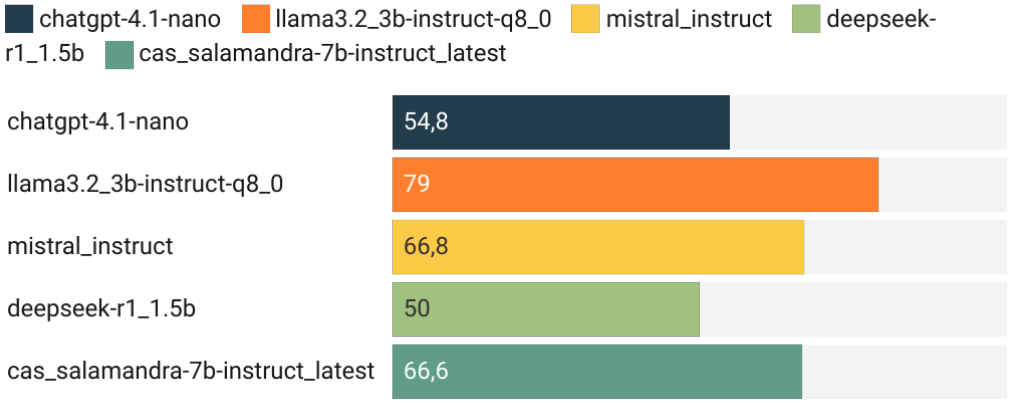
En primer lugar, esta dimensión mide cuán frecuentemente cada modelo identifica desequilibrio narrativo a favor de un actor en los textos evaluados (Gráfico 2). Se consideran indicadores como la presencia desigual de voces o la ausencia de contraposición. llama3.2:3b-instruct-q8_0 es el modelo que más tiende a detectar parcialidad en los textos (78.4%), seguido por cas/salamandra-7b (67.4%). Por el contrario, deepseek (38.8%) y mistral (45.2%) muestran menor reacción a estos elementos, lo que puede interpretarse como una postura más tolerante o menos perceptiva.

Gráfico 3: puntuación media de cada modelo en la dimensión de Ambigüedad Interpretativa



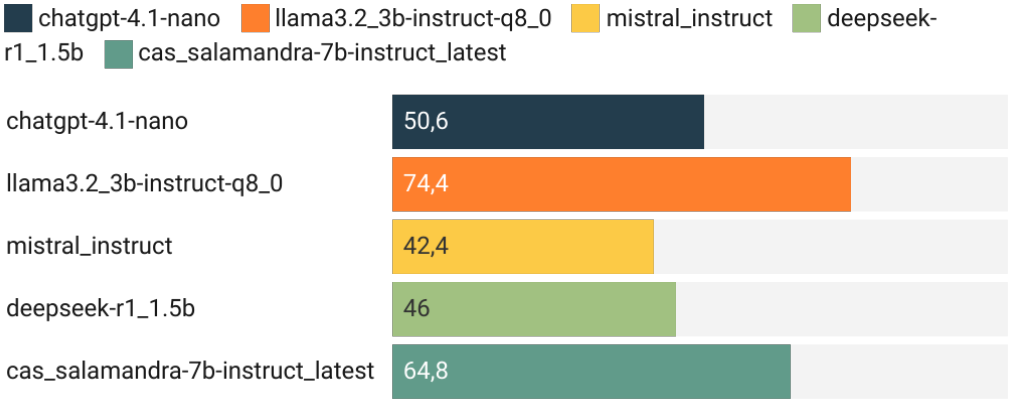
Evalúa hasta qué punto los modelos identifican vaguedad, imprecisión o ambivalencia en los textos (Gráfico 3). Detectar ambigüedad es relevante para interpretar el grado de apertura o indefinición narrativa. llama3.2 (80.8%) y cas/salamandra-7b (71.6%) destacan como los más sensibles, sugiriendo una alta capacidad (o predisposición) para detectar formulaciones abiertas o poco concluyentes. En cambio, chatgpt-4.1-nano (55%) parece menos atento a estas estructuras.

Gráfico 4: puntuación media de cada modelo en la dimensión de Dominancia Discursiva



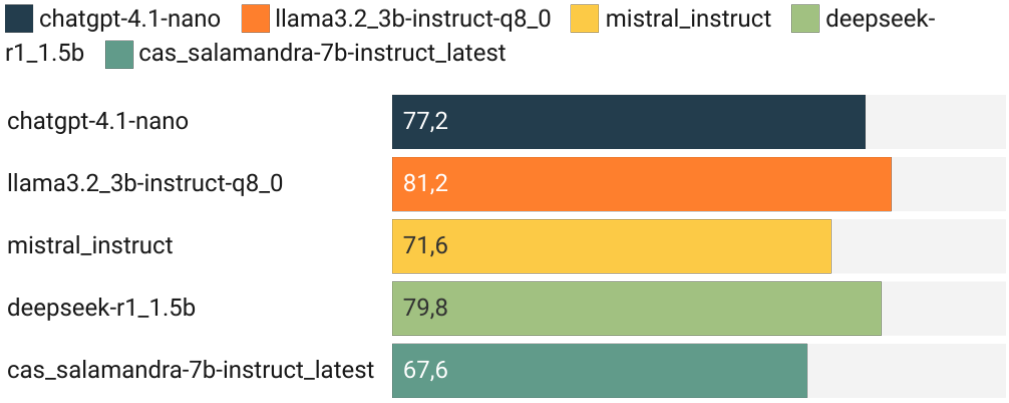
Esta dimensión refleja la frecuencia con que un modelo identifica que el texto está estructurado en torno a una única voz dominante (Gráfico 4). Es un indicador clave para evaluar pluralidad y equilibrio. Nuevamente, toma la delantera llama3.2 (79%) y destaca la alta puntuación de mistral (66.8%), siendo los modelos que más identifican dominancia en los textos, sugiriendo mayor atención al reparto de voz y a la diversidad narrativa. deepseek (50%) y chatgpt (54.8%) son más neutros, con menor nivel de alerta frente a discursos unilaterales.

Gráfico 5: puntuación media de cada modelo en la dimensión de Grado de Encuadre Ideológico



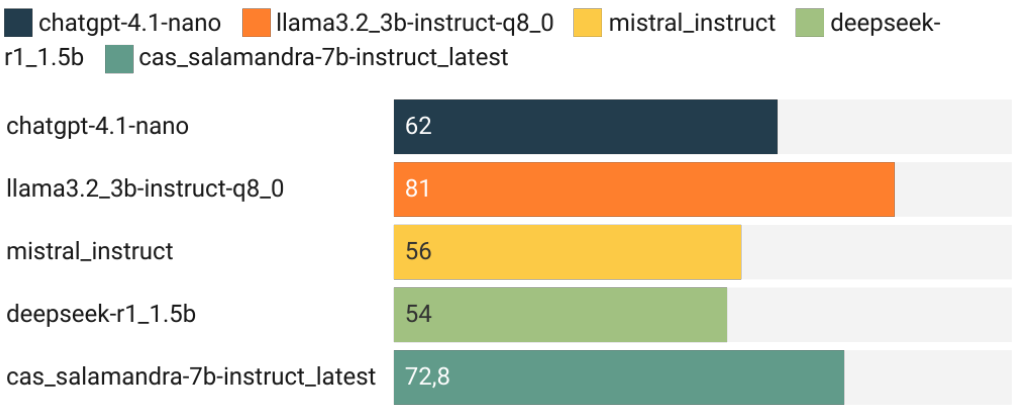
Aquí se analiza hasta qué punto los modelos identifican la presencia de marcos ideológicos explícitos o implícitos en los textos (Gráfico 5). Una puntuación alta no indica que el modelo sea ideológico, sino que percibe ideologización en lo que evalúa. llama3.2 (74.4%) y cas/salamandra-7b (64.8%) son los que más frecuentemente detectan encuadre ideológico. Esto puede deberse a una mayor sensibilidad ante expresiones marcadas políticamente. mistral (42.4%) y deepseek (46%) no lo identifican de forma tan firme, lo que puede reflejar un filtro más conservador en la detección de contenido ideologizado.

Gráfico 6: puntuación media de cada modelo en la dimensión de Claridad Estructural / Legibilidad



Esta dimensión analiza la capacidad del modelo para identificar una buena estructura y claridad lingüística en los textos evaluados (Gráfico 6). Una puntuación alta indica que el modelo considera que los textos son fáciles de seguir. Todos los modelos coinciden en evaluar favorablemente la claridad de los textos, teniendo en cuenta que han sido creados por ellos mismos. llama3.2 (81.2%) y deepseek (79.8%) son los que más frecuentemente consideran que el texto está bien estructurado, mientras que cas/salamandra-7b (67.6%) lo percibe con algo más de reserva.

Gráfico 7: puntuación media de cada modelo en la dimensión de Riesgo de Lectura Sesgada



La última de las dimensiones evalúa si el modelo detecta que un texto, aun sin ser explícitamente parcial, podría inducir a una interpretación polarizada por parte del lector (Gráfico 7). llama3.2 (81%) y cas/salamandra-7b (72.8%) son los más reactivos ante textos con alto potencial de interpretación sesgada. Este resultado sugiere una mayor alerta ante posibles efectos secundarios del discurso. Por el contrario, deepseek (54%) y mistral (56%) muestran una actitud más contenida en esta dimensión.

6.2.3. *Análisis de repetibilidad y consistencia entre ejecuciones*

Para comprobar la validez de los resultados, una de las preguntas centrales es si los *LLMs*, como agentes evaluadores, tienen consistencia interna. Esto es, si responden de forma estable ante los mismos estímulos, o su evaluación varía significativamente de una ejecución a otra. Para comprobarlo, cada uno de los modelos evaluadores fue sometido a cuatro ejecuciones independientes sobre la misma muestra de 50 noticias. Las condiciones de análisis fueron muy similares en cada iteración, permitiendo así comparar la estabilidad de sus puntuaciones.

Desviación estándar entre ejecuciones por cada dimensión en su escala de 0 a 5

Modelo	*Ambigüedad*	*Dominancia discursiva*	*Grado de encuadre ideológico*	*Parcialidad percivida*	*Riesgo de lectura sesgada*	*Claridad estructural / legibilidad*
ChatGPT	0.012	0.028	0.053	0.019	0.019	0.023
LLaMA	0.066	0.035	0.050	0.034	0.053	0.044
Mistral	0.084	0.066	0.025	0.050	0.066	0.028
Deepseek	0.084	0.229	0.263	0.171	0.189	0.141
Salamandra	0.107	0.066	0.091	0.048	0.060	0.050

Tabla 13: Desviación estándar entre ejecuciones por cada dimnesión en su escala de 0 a 5. Elaboración propia.

ChatGPT-4.1-nano presenta la menor variabilidad global, con desviaciones que oscilan entre 0.012 y 0.053. Su comportamiento es el más predecible y uniforme, especialmente en la dimensión de Ambigüedad interpretativa (0.012). LLaMA también mantiene una varianza muy baja, aunque algo más pronunciada en el encuadre ideológico (0.050) y riesgo de lectura sesgada (0.053). Mistral, por su parte, destaca por su estabilidad en el encuadre ideológico (0.025) y claridad (0.028), pero presenta cierta inestabilidad en ambigüedad y riesgo (0.084 y 0.066 respectivamente). DeepSeek, por el contrario, presenta una variabilidad más alta en todas las dimensiones, alcanzando desviaciones de hasta 0.263 en Grado de encuadre ideológico y 0.229 en Dominancia discursiva. Esto sugiere un comportamiento más inconsistente o sensible a pequeñas variaciones en entrada o contexto, aunque igualmente despreciables. Finalmente, Salamandra presenta un perfil intermedio. Aunque más estable que DeepSeek, muestra mayor variabilidad que ChatGPT, LLaMA o Mistral, especialmente en Ambigüedad (0.107) y Encuadre ideológico (0.091).

En conjunto, las desviaciones son notablemente bajas, lo que indica un comportamiento altamente estable entre ejecuciones. Este resultado refuerza la validez de los análisis automatizados y aumenta la credibilidad de los resultados obtenidos, al disminuir la influencia de la aleatoriedad residual típica de los *LLMs*. Esta consistencia permite considerar que las diferencias observadas entre modelos pueden ser atribuibles a sus características internas, y no al azar. DeepSeek y Salamandra, al mostrar mayor variación, podrían requerir un mayor número de ejecuciones para estabilizar sus métricas en futuros estudios.

6.2.4. Correlaciones entre dimensiones

Matriz de correlación de Pearson entre las seis dimensiones analizadas por los LLMs

	Parcialidad percibida	***Ambigüedad interpretativa***	***Dominancia discursiva***	***Grado de encuadre ideológico***	***Claridad estructural / legibilidad***	***Riesgo de lectura sesgada***
Parcialidad percibida	1.00	0.27	0.33	0.46	-0.03	0.45
Ambigüedad interpretativa	0.27	1.00	0.22	0.26	0.04	0.31
Dominancia discursiva	0.33	0.22	1.00	0.31	0.08	0.29
Grado de encuadre ideológico	0.46	0.26	0.31	1.00	0.08	0.35
Claridad estructural / legibilidad	-0.03	0.04	0.08	0.08	1.00	0.04
Riesgo de lectura sesgada	0.45	0.31	0.29	0.35	0.04	1.00

Tabla 14: Matriz de correlación de Pearson entre las seis dimensiones analizadas por los LLMs. Elaboración propia.

Con el fin de examinar las posibles relaciones internas entre los distintos aspectos evaluados en las noticias generadas por los *LLMs*, se ha calculado la matriz de correlación de Pearson entre las seis dimensiones que conforman el índice compuesto (Tabla 14).

Los resultados muestran que existen correlaciones moderadas a fuertes entre ciertas dimensiones, especialmente aquellas vinculadas con la carga ideológica del texto. La parcialidad percibida presenta asociaciones destacadas con el grado de encuadre ideológico ($r = 0.46$) y con el riesgo de lectura sesgada ($r = 0.45$), lo que sugiere que cuando los *LLMs* perciben una orientación discursiva parcial, también identifican encuadre y potencial de interpretación sesgada. Asimismo, se observa una correlación moderada entre la dominancia discursiva y el grado de encuadre ideológico ($r = 0.31$), así como entre esta misma dimensión y el riesgo de lectura sesgada ($r = 0.29$). Estos resultados refuerzan la hipótesis de que un discurso donde predominan ciertas voces o perspectivas también tiende a estar ideológicamente enmarcado y con menor pluralidad interpretativa.

Por el contrario, la dimensión de claridad estructural / legibilidad no presenta correlaciones significativas con ninguna de las otras variables (máxima $r = 0.08$), e incluso aparece débilmente correlacionada de forma negativa con la parcialidad percibida ($r = –0.03$). Esto sugiere que la claridad formal del texto evaluado opera como una dimensión estructural relativamente autónoma, no influida por la orientación ideológica ni por el contenido narrativo.

7. CONCLUSIONES

Este trabajo confirma que los *LLMs* manifiestan sesgos ideológicos y geopolíticos cuando se enfrentan a tareas de generación y evaluación de contenidos construidos sobre escenarios ambiguos o complejos. A través de un diseño experimental que articuló generación narrativa, análisis estructurado y evaluación cruzada, se validan tanto el objetivo general como los tres objetivos específicos formulados.

En relación con la hipótesis general, los resultados muestran con claridad que los LLMs no operan desde una neutralidad. Incluso en ausencia de directrices valorativas explícitas, tienden a completar las narrativas siguiendo que reflejan una orientación ideológica. Esta hipótesis encuentra apoyo en múltiples *outputs*: en una noticia generada por ChatGPT a partir de un escenario sobre tensiones en Europa del Este (*chatGPT_scenario_001*), se afirma directamente que *"la agresión de la República de Zorvia ha desestabilizado la región"*, a pesar de que el escenario original no atribuía responsabilidad a ningún actor.

Respecto a la **hipótesis H1**, se constata que los modelos reproducen encuadres coherentes con los valores dominantes de sus contextos de entrenamiento. Esto se aprecia en la reiteración de fórmulas retóricas como *"defensa de la soberanía"* para referirse a países alineados con potencias occidentales o en la omisión de problematizaciones sobre actores aliados. Múltiples modelos adoptan de forma automática marcos como *"libertad de navegación"* o "*violación de normativas internacionales"*.

La **hipótesis H2** también se ve respaldada: los textos generados presentan estructuras narrativas recurrentes que revelan un posicionamiento discursivo reconocible. Por ejemplo, en el corpus de DeepSeek, varias noticias describen maniobras chinas como *"provocaciones armadas"* mientras se enmarca a Estados Unidos como garante de la estabilidad. Este tipo de framing se repite incluso cuando el escenario no establece de forma clara las acciones militares de los actores implicados.

En cambio, la **hipótesis H3** presenta resultados más ambivalentes. Aunque algunos modelos como ChatGPT ofrecieron análisis razonablemente estructurados, otros como DeepSeek mostraron puntuaciones incoherentes o contradictorias entre variables clave.

En cuanto a los objetivos específicos, se han alcanzado satisfactoriamente:

- Se logró la generación de escenarios geopolíticos ambiguos **(O1)**.
- Se realizó un análisis de los sesgos presentes en más de 250 noticias **(O2)**.
- Se implementó un sistema de evaluación cruzada entre modelos **(O3)**.

Los resultados obtenidos permiten afirmar que los *LLMs* no solo producen texto, sino también sentido. Actúan como agentes de estructuración simbólica, reproduciendo y estabilizando interpretaciones del mundo incluso cuando la entrada ha sido cuidadosamente neutralizada. Así, estas conclusiones dejan claro que el sesgo en los *LLMs* no es un efecto marginal, sino un fenómeno estructural.

8. DISCUSIÓN

Los hallazgos resultantes del presente trabajo de investigación deben entenderse dentro de una serie de limitaciones y retos metodológicos, técnicos y estructurales que es necesario discutir para valorar el alcance real del estudio y plantear vías de mejora futura.

Una de las principales limitaciones detectadas es la ausencia de evaluación humana directa en la puntuación de los textos generados. Aunque la rúbrica construida permite una valoración estructurada, replicable y sistemática, su aplicación exclusiva por parte de otros *LLMs* introduce sesgos metodológicos importantes, especialmente en tareas que requieren juicio crítico o interpretación ideológica. El hecho de que los propios modelos evalúen textos generados por ellos mismos o por modelos con arquitecturas similares compromete, en parte, la independencia del análisis. Además, incluir evaluadores humanos expertos permitiría estudiar la capacidad de estos para detectar sesgos implícitos.

En otro respecto, la inestabilidad de los *outputs* generados, especialmente en modelos como DeepSeek o LLaMA. En varios casos, la puntuación de una misma noticia varió significativamente entre evaluaciones independientes, y las diferencias entre *outputs* generados a partir de un mismo escenario fueron suficientemente amplias como para comprometer comparaciones simples. Esto afecta directamente al testeo de la hipótesis H3 sobre la coherencia interna de los modelos en tareas de evaluación cruzada. Una posible solución sería aumentar el número de iteraciones por escenario y modelo, permitiendo así una normalización estadística más robusta. Trabajar con 5 o más muestras por escenario y por modelo ayudaría a reducir la influencia del azar y permitiría obtener medidas más estables de sesgo discursivo.

Además, aunque no menos importante, el estudio ha estado condicionado por limitaciones materiales importantes. Por un lado, la ejecución local de modelos como Mistral o DeepSeek ha requerido una gestión manual de entornos que, aunque viable, no garantiza condiciones idénticas de inferencia en todos los casos. Por otro lado, la evaluación de ChatGPT a través de *API* (*Application Programming Interface*) ha dificultado la replicación de análisis y ha introducido incertidumbre metodológica. Contar con una infraestructura más potente permitiría utilizar modelos con mayor número de parámetros y contrastar si el sesgo disminuye, se estabiliza o simplemente se enmascara con mayor sofisticación lingüística. Asimismo, disponer de financiación para el uso de *APIs* comerciales abriría la puerta al análisis sistemático de modelos propietarios como GPT-4, Claude o Gemini, cuya influencia real en el ecosistema comunicativo es mucho mayor que la de los modelos *open-source* actualmente accesibles.

Por otro lado, el código se ha constituido como uno de los pilares técnicos del trabajo, materializado en el desarrollo de *biasLens*. Aunque su estructura actual es robusta y ha permitido gestionar de forma sistemática el flujo completo del experimento, la herramienta podría beneficiarse de una revisión más exhaustiva de su arquitectura, la automatización de procesos, la validación formal de entradas/salidas y una mayor modularización. Esto permitiría no solo mejorar la eficiencia del proceso, sino facilitar la replicación del estudio por otros equipos de investigación. Una versión más pulida de *biasLens*, acompañada de documentación ampliada y de ejemplos reproducibles, podría convertirse en una herramienta

abierta útil para el análisis crítico de sesgos discursivos en *LLMs* desde disciplinas como la comunicación, el periodismo o la ciencia política.

De esta manera, y a pesar de estar aún en fase de desarrollo, *biasLens* constituye una aportación con potencial de continuidad. Entre las líneas futuras más prometedoras se encuentra la incorporación de evaluadores humanos al sistema, la creación de una interfaz visual accesible para investigadores no técnicos y la integración de métricas discursivas más complejas. De este modo, la herramienta podría evolucionar hacia una plataforma híbrida capaz de combinar evaluación automatizada y análisis cualitativo, abriendo así nuevas posibilidades para la auditoría crítica de tecnologías generativas en el espacio público.

Por último, es importante señalar que el propio concepto de sesgo ideológico no es unívoco ni está cerrado. La clasificación de ciertos enunciados como "sesgados" depende en parte de marcos teóricos y epistemológicos que, aunque fundados en la teoría del *Framing*, la hegemonía o la *Agenda Setting*, siguen siendo interpretativos. Por ello, el trabajo no pretende establecer un patrón universal de sesgo, sino proponer una metodología razonada y aplicable para identificar estructuras narrativas que orientan la interpretación de la realidad. Este enfoque no elimina el problema del sesgo, pero permite al menos tratarlo desde una posición crítica, informada y replicable. Lejos de sugerir que el sesgo puede eliminarse completamente, este estudio apunta a la necesidad de alfabetización crítica en el uso de tecnologías generativas, especialmente cuando estas son utilizadas en contextos informativos o institucionales.

9. BIBLIOGRAFÍA

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*. http://arxiv.org/abs/2005.14165

Entman, R. M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, *43*(4), 51-58.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2023). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, *50*(3). http://arxiv.org/abs/2309.00770

Katz, E., Blumler, J. G., & Gurevitch, M. (1973). Uses and gratifications research. *Public Opinion Quarterly*, *37*(4), 509-523. https://doi.org/https://doi.org/10.1086/268109

Lin, X., & Li, L. (2024). *Implicit Bias in LLMs: A Survey*. https://arxiv.org/abs/2503.02776

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, *36*(2), 176-187. https://doi.org/https://doi.org/10.1086/267990

McLuhan, M. (2013). *Understanding media: the extensions of a man* (3.ª ed.). GINGKO PRESS Inc.

Noelle-Neumann, E. (1974). The Spiral of Silence: A Theory of Public Opinion. *Journal of Communication*, *24*(2), 43-51. https://doi.org/https://doi.org/10.1111/j.1460-2466.1974.tb00367.x

Ranjan, R., Gupta, S., & Narayan Singh, S. (2024). *A Comprehensive Survey of Bias in LLMs: Current Landscape and Future Directions*. https://arxiv.org/pdf/2409.16430

Semetko, H. A., & Valkenburg, P. M. (2000). Framing European Politics: A Content Analysis of Press and Television News. *Journal of Communication*, *50*(2), 93-109. https://doi.org/https://doi.org/10.1111/j.1460-2466.2000.tb02843.x

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, diciembre). Attention Is All You Need. *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. http://arxiv.org/abs/1706.03762

10. GLOSARIO

[i] **LLM**: modelo de lenguaje de gran tamaño entrenado con enormes cantidades de parámetros para generar, traducir o analizar lenguaje natural.

[ii] ***Prompt***: entrada de texto que se le da a un modelo de lenguaje para que genere una respuesta coherente y contextualizada. Dentro de el se distinguen el *user_prompt* (input del usuario) y *system_prompt* (contexto/instrucciones).

[iii] **Parámetros**: valores ajustables en una red neuronal que determinan su comportamiento; los *LLMs* pueden tener millones, billones o incluso trillones. Los forman una pregunta y una respuesta.

[iv] **Q8**: técnica de cuantización que reduce el tamaño de los modelos de IA representando los valores en 8 bits, manteniendo buen rendimiento con menor consumo.

[v] **Instruct**: tipo de modelo de IA entrenado para seguir instrucciones específicas dadas por el usuario, generando respuestas más útiles y orientadas a tareas.

[vii] ***Open-source***: *software* cuyo código fuente está disponible públicamente para ser modificado y distribuido libremente.

[viii] **Chat (tipo de modelo)**: modelo conversacional diseñado para enriquecer las respuestas más allá del contexto inmediato.

[ix] ***Reinforced Learning (training schema)***: técnica de aprendizaje automático donde un agente aprende mediante recompensas y penalizaciones para tomar decisiones optimas.

[x] ***Mixture of Experts (training schema)***: arquitectura de red neuronal donde diferentes submodelos (expertos) se activan según la tarea, optimizando recursos y precisión.

[xi] ***Script***: conjunto de instrucciones o código que automatiza tareas o procesos en un entorno informático.

[xii] ***Assistant*** **(OpenAI)**: método de entrada actual a la *API* de Openai. Actualmente en proceso de sustitución.

[xiii] ***Fine-tuning (training schema)***: proceso de ajustar un modelo preentrenado con nuevos datos específicos para mejorar su rendimiento en tareas concretas.

[xiv] ***In-context-learning (training schema)***: método donde el modelo generaliza o responde correctamente solo a partir del contexto proporcionado en el mismo *prompt*, sin entrenamiento adicional.

[xv] ***Few-shot & zero-shot (training-schema)***: enfoques donde el modelo realiza tareas con pocos ejemplos (*few-shot*) o ninguno (*zero-shot*) durante la inferencia (entrenamiento), gracias a su comprensión general previa.