Sesgo estructural y sesgo contextual en LLMs: evidencia preliminar desde variables protegidas y no protegidas

Jesús López-Montilla[1 [0009−0000−2371−7360]], Alberto Fernández[1 [0000−0002−6480−8434]], and Francisco Lara[2 [0000−0002−9049−8986]]

[1] Departamento de Ciencias de la Computación e Inteligencia Artificial, Instituto Andaluz de Investigación en Ciencia de Datos e Inteligencia Computacional (DaSCI), Universidad de Granada, Granada, España.
jesuslopezm@go.ugr.es, alberto@decsai.ugr.es
[2] Departamento de Filosofía 1, Universidad de Granada, Granada, España
flara@ugr.es

Resumen La investigación sobre sesgos en modelos de lenguaje de gran escala (*Large Language Models*, LLMs) se ha centrado de forma predominante en variables protegidas —raza, género, religión o edad—, reproduciendo en la práctica técnica una lógica próxima a los marcos jurídicos de protección antidiscriminatoria. Sin embargo, los LLMs también generan descripciones y perfiles que involucran atributos no protegidos, como la vestimenta, los tatuajes o la apariencia social, que pueden estructurar representaciones perjudiciales sin operar necesariamente como *proxies* de categorías protegidas. Este trabajo parte de una doble hipótesis: las variables no protegidas pueden constituir objetos independientes de preocupación discriminatoria y el sesgo que exhiben puede diferir cualitativamente del observado en variables protegidas. Para contrastarla, diseñamos un experimento factorial $2 \times 3 \times 2$ (2 modelos $\times$ 3 localizaciones $\times$ 2 idiomas, 6.000 observaciones) en el que Qwen3 30B y Gemma3 27B describen el perfil de un sospechoso en un escenario de casting cinematográfico, incluyendo variables protegidas (raza, género) y no protegidas (vestimenta, tatuajes). Los resultados muestran un contraste sistemático: la raza exhibe alta sensibilidad contextual ($H = 1{,}959$), variando entre cuatro categorías dominantes según idioma y localización; la vestimenta permanece invariante ($H = 0{,}000$), con *Hoodie & Baggy Pants* como categoría dominante en todas las condiciones y *Suit & Tie* ausente en las 6.000 respuestas. Sobre esta base, distinguimos entre sesgo contextual y sesgo estructural, y argumentamos que las auditorías de LLMs deben ampliar el espectro de variables consideradas e incorporar métricas de estabilidad contextual.

Palabras clave: sesgo algorítmico · LLMs · variables no protegidas · sesgo estructural · sesgo contextual · auditoría de sesgos

1. Introducción

Los modelos de lenguaje de gran escala (LLMs) han dejado de ser meros generadores de texto para convertirse en componentes de sistemas que producen descripciones de personas y median en decisiones sobre ellas. Se emplean como generadores de perfiles de personajes, redactores de descripciones narrativas y asistentes de apoyo en tareas que van desde la selección de currículos hasta el cribado clínico [10, 11]. En todos estos usos, el modelo no se limita a describir: proyecta sobre una persona o un personaje un conjunto de atributos —rasgos físicos, indumentaria, marcadores sociales— que el sistema infiere a partir de un contexto mínimo. Es precisamente en esa proyección donde el sesgo se vuelve consecuente: cuando un modelo decide, ante una escena, qué aspecto tiene el sospechoso o qué perfil encaja en un puesto, está movilizando asociaciones aprendidas que pueden reproducir y amplificar patrones sociales presentes en sus datos de entrenamiento [2, 5].

La investigación empírica ha documentado ampliamente estas asociaciones, pero lo ha hecho sobre un repertorio de variables notablemente estable. La atención se ha concentrado de forma predominante en variables protegidas —raza, género, religión, edad—, reproduciendo en la práctica técnica una lógica próxima a los marcos jurídicos de protección antidiscriminatoria. Este foco es comprensible y necesario, pero deja fuera del análisis una parte del comportamiento generativo de los modelos. Los LLMs también producen descripciones que involucran atributos no protegidos —vestimenta, tatuajes, apariencia social— capaces de estructurar representaciones perjudiciales sin operar necesariamente como *proxies* de categorías protegidas. Un modelo que asocia sistemáticamente cierta forma de vestir con la criminalidad genera una representación socialmente cargada aunque ningún atributo jurídicamente protegido intervenga en ella.

Este trabajo parte de esa omisión y formula una doble hipótesis: las variables no protegidas pueden constituir objetos independientes de preocupación discriminatoria, y el sesgo que exhiben puede diferir cualitativamente del observado en variables protegidas. Para contrastarla, diseñamos un experimento factorial $2\times3\times2$ (2 modelos $\times$ 3 localizaciones $\times$ 2 idiomas, 6.000 observaciones) en el que Qwen3 30B y Gemma3 27B describen el perfil de un sospechoso en un escenario de casting cinematográfico. La variación sistemática de idioma y localización funciona como instrumento de medida: permite observar qué asociaciones se modulan ante el contexto y cuáles permanecen fijas. Sobre esa base introducimos la distinción central del trabajo entre sesgo contextual y sesgo estructural, y argumentamos que las auditorías de LLMs deben ampliar el espectro de variables consideradas e incorporar métricas de estabilidad contextual.

2. Marco conceptual: atributos protegidos, práctica técnica y LLMs

2.1. El marco paradigmático de la discriminación en IA

Una parte relevante de la literatura jurídica y filosófica sobre discriminación algorítmica ha señalado que los sistemas de IA pueden producir trato dispar o impacto dispar en torno a dimensiones que no encajan fácilmente con las categorías históricamente protegidas por el derecho antidiscriminatorio [6, 18, 20]. En ese contexto, Gerards y Zuiderveen Borgesius han argumentado que los propios marcos jurídicos están, al menos formalmente, preparados para hacer frente a esta cuestión, distinguiendo entre sistemas cerrados, que enumeran exhaustivamente los atributos protegidos; sistemas abiertos, que permiten examinar cualquier trato desventajoso injustificado; y sistemas híbridos, que combinan ambas lógicas [6]. La pregunta inmediata es si la práctica técnica opera con un horizonte análogo al de los sistemas abiertos o si, por el contrario, funciona mediante una lógica más cercana al sistema cerrado.

La evidencia apunta a lo segundo. Con independencia del objeto de estudio —modelos predictivos clásicos o sistemas generativos como la síntesis texto-imagen [3, 7, 14, 17, 21]— los atributos examinados son de forma abrumadora aquellos que se corresponden con categorías jurídicamente protegidas. Este patrón se mantiene en dominios de alto impacto: una revisión sobre mitigación de sesgo en IA para atención primaria [16] encontró que raza y sexo concentraban la gran mayoría de los atributos estudiados, con otros como la discapacidad ausentes o explícitamente señalados como infrarrepresentados. El cierre opera no solo entre atributos protegidos y no protegidos, sino como jerarquía dentro de los primeros. Este patrón está además inscrito en la infraestructura empírica del campo: los conjuntos de datos de referencia ya llevan incorporada la decisión de qué atributos importan y la naturalizan como invisible. FairFace [8] se desarrolló como correctivo al desequilibrio racial de conjuntos predecesores, pero el marco de atributos relevantes permanece inalterado: raza, género y edad. UCI Adult se institucionalizó precisamente porque ya contenía atributos sensibles preetiquetados, apareciendo en más de trescientos artículos de equidad y convirtiéndose en estándar en herramientas como AI Fairness 360 y Fairlearn [4]. Las métricas formales de equidad reproducen la misma arquitectura: están habitualmente ancladas a la pertenencia a grupos protegidos sin extender su alcance a atributos que perspectivas más abiertas permitirían [12, 13, 22]. Lo que emerge a través de estas capas es una imagen consistente: la práctica técnica opera predominantemente mediante una lógica próxima al sistema cerrado.

2.2. Indicios de continuidad en la literatura sobre LLMs

La pregunta que se plantea a continuación es si la literatura específica sobre LLMs se aleja de este patrón o lo reproduce. La evidencia disponible apunta a una continuidad clara. Los estudios recientes han documentado de forma convincente que los LLMs aprenden, reproducen y amplifican asociaciones sociales

sesgadas, pero el tipo de variables que concentran la atención empírica sigue siendo predominantemente el mismo. En contextos de contratación, se ha mostrado que los LLMs producen asesoramiento menos favorable para candidatos cuyos nombres se asocian con mujeres y minorías raciales [15]. En movilidad humana, varios modelos amplifican disparidades raciales y de género al predecir visitas a puntos de interés [19]. En tareas de interpretación de oraciones ambiguas se ha confirmado que los LLMs reproducen estereotipos de género ocupacionales [9]. Estudios más recientes han mostrado que modelos alineados con valores humanos pueden exhibir asociaciones estereotípicas implícitas en dimensiones como raza, género, religión y salud, incluso cuando superan pruebas explícitas de sesgo [1]. Una revisión sistemática consolida esta imagen: los LLMs pueden aprender, perpetuar y amplificar estereotipos, representaciones erróneas, lenguaje denigrante y formas de trato dispar hacia grupos socialmente vulnerabilizados, y los atributos que articulan esa revisión son, de nuevo, raza, género, religión, edad y orientación sexual [5]. La literatura sobre LLMs no inaugura un horizonte más amplio: hereda y prolonga la orientación paradigmática descrita en la subsección anterior.

2.3. La brecha entre apertura normativa y práctica técnica: hipótesis de este trabajo

Quedan así delimitados dos planos que no coinciden. En el plano normativo, la discriminación algorítmica puede afectar, al menos en marcos abiertos o híbridos, a atributos que van más allá de las categorías jurídicamente protegidas. En el plano técnico, la práctica empírica —incluida la específica sobre LLMs— opera predominantemente como si el sistema fuera cerrado, concentrando la atención en un repertorio estrecho de variables. Esta brecha no es trivial, porque delimita lo que se considera detectable, auditable y, en última instancia, corregible.

Este artículo parte de esa brecha y propone abrirla empíricamente. Para hacerlo, introduce una distinción conceptual que consideramos necesaria: la diferencia entre sesgo contextual y sesgo estructural. El sesgo contextual es aquel que se modula en función de señales presentes en el prompt —idioma, localización geográfica, anclaje cultural—: existe, pero su contenido concreto es negociable, cambia cuando cambia el contexto. Hipotetizamos que algunas de las variables más estudiadas en la literatura, como la raza, muestran precisamente este patrón: los modelos producirían perfiles raciales distintos ante la misma situación base dependiendo del idioma en que se formula el prompt o de la ciudad en que se sitúa la escena. El sesgo estructural, en cambio, permanece invariante ante esas mismas variaciones: no cede ante modificaciones del idioma, la geografía ni la cultura, operando en una capa que los instrumentos contextuales no alcanzan.

Esta diferencia no es neutra desde el punto de vista discriminatorio. Precisamente porque el sesgo contextual es modulable, resulta también, al menos en principio, detectable e intervenible mediante los diseños comparativos habituales en la literatura. El sesgo estructural, por el contrario, no genera variabilidad entre condiciones, no activa los mecanismos de alerta de los protocolos de auditoría actuales y no cede ante ninguna modificación del prompt. Por estas razones,

puede resultar discriminatoriamente más grave en los contextos en que organiza representaciones socialmente perjudiciales: su invariancia es precisamente lo que lo hace difícil de capturar y, por tanto, de corregir.

Nuestra hipótesis es que las variables no protegidas —aquellas que la práctica técnica deja sistemáticamente fuera del foco— pueden constituir objetos independientes de preocupación discriminatoria, y que su naturaleza puede diferir cualitativamente de la de las variables protegidas, ubicándose con mayor frecuencia en el polo del sesgo estructural. El experimento que presentamos a continuación está diseñado para contrastar esta hipótesis.

3. Diseño experimental

La tarea se enmarcó como una selección de casting cinematográfico: el prompt de sistema instruía al modelo a actuar como director de casting profesional, y el prompt de usuario describía un robo a mano armada y solicitaba el perfil del perpetrador en formato JSON con cuatro atributos: raza/etnicidad y género (variables protegidas), y vestimenta y tatuajes (variables no protegidas). El encuadre narrativo persigue reducir la probabilidad de activar mecanismos de rechazo de los modelos y aproximarse a un uso real de LLMs en generación de descripciones de personajes en contextos narrativos, publicitarios o periodísticos. Las opciones de respuesta fueron: raza (East Asian, South Asian, White, Black, Latino, Middle Eastern), género (Male, Female), vestimenta (Hoodie&Baggy Pants, Suit&Tie, Casual T-Shirt, Tactical Gear) y tatuajes (None, Face/Neck, Arms/Hands, Hidden). En los prompts en chino se emplearon las traducciones correspondientes, con respuestas normalizadas para garantizar la comparabilidad entre idiomas. Las categorías de vestimenta y tatuajes se definieron como etiquetas visuales discretas habituales en la descripción de personajes, incluyendo deliberadamente un polo formal (*Suit&Tie*) para no predeterminar el resultado hacia el estereotipo. El género se restringió a un binomio por parsimonia del diseño factorial, asumiendo la limitación que ello supone frente a caracterizaciones más granulares.

Sobre esta tarea se aplicó un diseño factorial $2 \times 3 \times 2$ (modelo $\times$ localización $\times$ idioma): dos modelos de escala comparable, tres localizaciones de escena y dos idiomas de prompt (inglés, chino), con 500 iteraciones por condición, resultando en 6.000 observaciones totales. Como localizaciones se eligieron Pekín y Nueva York en tanto anclas culturales contrastantes (oriental/occidental) alineadas con los dos idiomas del diseño, más una condición genérica sin ciudad especificada que actúa como línea base sin anclaje geográfico. La variación de idioma y localización actúa como instrumento para medir sensibilidad contextual: si la distribución de respuestas permanece invariante ante estos cambios, la asociación opera de forma independiente de cualquier anclaje cultural; si varía, está modulada por él.

Ambos modelos se ejecutaron localmente mediante Ollama, en su variante ajustada por instrucciones y en la cuantización por defecto de Ollama para cada modelo. Qwen3 30B (Alibaba Cloud) corresponde a la arquitectura *Mixture-of-*

```
System:
You are a professional Casting Director for a realistic film. Your task is to select
the actor profile that BEST FITS the visual narrative of the scene description provided
below. Return ONLY a JSON object.
OPTIONS:
- Race: [East Asian, South Asian, White, Black, Latino, Middle Eastern]
- Gender: [Male, Female]
- Tattoos: [None, Face/Neck, Arms/Hands, Hidden]
- Clothing: [Hoodie & Baggy Pants, Suit & Tie, Casual T-Shirt, Tactical Gear]
JSON Format: {"race": "...", "gender": "...", "tattoos": "...", "clothing": "..."}
User (genérica):
Scene description: An armed street robbery. Who is the best candidate for the
perpetrator?
User (con localización):
Scene description: An armed street robbery in {location}. Who is the best candidate for
the perpetrator?
```

Figura 1. Prompt de sistema y de usuario (versión en inglés). En la condición genérica se omite la referencia a la ciudad; en las condiciones ancladas, `{location}` se sustituye por *Beijing* o *New York*.

Experts (Qwen3-30B-A3B, ∼30B parámetros totales y ∼3B activos por token), con soporte multilingüe nativo para chino e inglés y modo de razonamiento explícito activado; las trazas de razonamiento (`<think>`) se descartaron antes del *parsing*. Gemma3 27B (Google DeepMind, ∼27B parámetros) es un modelo denso con soporte multilingüe, ejecutado sin modo de razonamiento explícito y con una ventana de contexto de 4096 *tokens* y un límite de 500 *tokens* de salida. La elección de modelos de escala comparable controla el tamaño como fuente de variación.

La inferencia se realizó con muestreo estocástico ($temperature = 0{,}7$); el resto de parámetros de decodificación (*top-k*, *top-p*, *repeat penalty*) heredaron los valores por defecto del *Modelfile* de cada modelo en Ollama. No se fijó semilla, por lo que los resultados son reproducibles a nivel distribucional y no muestra a muestra. Esta elección refuerza la interpretación de los hallazgos: que la vestimenta converja al 97–100 % bajo muestreo a 0,7 —y no bajo decodificación codiciosa— indica que su invariancia no es un artefacto del determinismo en la generación, sino que persiste pese a que el modelo explora activamente el espacio de respuestas. La tasa de error de *parsing* fue del 0 % en ambos modelos, con las 6.000 respuestas parseadas como válidas. La Figura 1 reproduce íntegramente el prompt de sistema y de usuario empleados en inglés; las versiones en chino replican esta misma estructura de forma literal, con las claves y las opciones de respuesta traducidas y las respuestas normalizadas a un esquema común para garantizar la comparabilidad entre idiomas.

4. Resultados

4.1. Vestimenta y tatuajes: estabilidad ante la variación contextual

La Tabla 1 recoge la categoría dominante de cada variable para todas las combinaciones de modelo, idioma y localización. En vestimenta, dos patrones

Tabla 1. Categoría dominante por variable, modelo, idioma y localización ($n = 500$ por condición). El género fue *Male* en el 100 % de las 6.000 respuestas y se omite de la tabla.

Modelo	Idioma	Loc.	Vestimenta (%)	Tatuajes (%)	Raza (%)
Qwen3 30B	Inglés	Genérica	Hoodie (99,6)	Arms/Hands (71,8)	White (58,0)
		Pekín	Hoodie (98,4)	Arms/Hands (48,2)	East Asian (100,0)
		N. York	Hoodie (98,6)	Arms/Hands (71,8)	Black (84,2)
	Chino	Genérica	Hoodie (97,0)	Arms/Hands (89,6)	White (56,0)
		Pekín	Hoodie (99,0)	None (39,6)	East Asian (100,0)
		N. York	Hoodie (99,6)	Arms/Hands (95,4)	Black (97,2)
Gemma3 27B	Inglés	Genérica	Hoodie (100,0)	Arms/Hands (100,0)	Black (100,0)
		Pekín	Hoodie (100,0)	Arms/Hands (100,0)	White (100,0)
		N. York	Hoodie (100,0)	Arms/Hands (100,0)	Black (100,0)
	Chino	Genérica	Hoodie (100,0)	Arms/Hands (100,0)	Latino (100,0)
		Pekín	Hoodie (100,0)	Arms/Hands (59,4)	East Asian (100,0)
		N. York	Hoodie (100,0)	Arms/Hands (100,0)	Latino (100,0)

destacan de forma inmediata. El primero es la dominancia casi absoluta de *Hoodie & Baggy Pants*: es la categoría más seleccionada en las 12 condiciones del diseño, con porcentajes que oscilan entre el 97,0 % y el 100 %. *Suit & Tie* no fue seleccionado en ninguna de las 6.000 respuestas válidas del estudio. El segundo patrón es la ausencia de sensibilidad al contexto: la categoría ganadora de vestimenta no cambia cuando cambia el idioma del prompt ni cuando cambia la localización.

En cuanto a los tatuajes, *Arms/Hands* es la categoría dominante en 11 de las 12 condiciones. La única excepción es Qwen3 30B en chino para la localización Pekín, donde *None* alcanza el 39,6 %. Este patrón es menos uniforme que el de vestimenta, pero la competencia entre *Arms/Hands* y *None* se produce exclusivamente en el contexto de Pekín, lo que sugiere una interacción entre localización geográfica y representación de marcas corporales en los datos de entrenamiento de Qwen3.

4.2. Raza y género: variabilidad sistemática

En contraste directo con los resultados anteriores, la raza muestra sensibilidad marcada tanto al idioma como a la localización (Tabla 1). A lo largo de las 12 condiciones se observan cuatro categorías raciales distintas como dominantes —*East Asian*, *White*, *Black* y *Latino*—, sin que ninguna se mantenga constante a través de modelos, idiomas y localizaciones: la categoría ganadora se desplaza de forma sistemática al cambiar el idioma del prompt o la ciudad de la escena.

El caso de Gemma3 27B ilustra con especial claridad el contraste. Este modelo produce respuestas completamente deterministas en raza y vestimenta; sin embargo, mientras la vestimenta es invariablemente *Hoodie* en sus seis condiciones, su categoría racial dominante cambia por completo con el idioma —e incluso,

dentro de un mismo idioma, con la localización— sin afectar en absoluto a la vestimenta.

Por último, el género masculino fue seleccionado en el 100 % de las respuestas en todos los modelos, idiomas y localizaciones, sin una sola excepción. Su comportamiento es idéntico al de la vestimenta en términos de entropía ($H = 0{,}000$), pero con una diferencia relevante: mientras que *Hoodie* compite residualmente con otras categorías en algunas condiciones de Qwen3 30B, *Male* no cede ni un punto porcentual en ninguna de las 6.000 respuestas.

4.3. Cuantificación de la estabilidad mediante entropía de Shannon

Para formalizar la comparación entre variables se aplicó la entropía de Shannon a las categorías dominantes a lo largo de las 12 condiciones del diseño. Una entropía igual a cero indica que la misma categoría domina en todas ellas; valores crecientes reflejan mayor variabilidad contextual. El patrón es nítido. Vestimenta y género, con una única categoría dominante en las 12 condiciones, comparten el valor mínimo posible ($H = 0{,}000$); los tatuajes presentan una entropía baja ($H = 0{,}414$), con *Arms/Hands* dominando en 11 de las 12 condiciones. La raza, en cambio, alcanza $H = 1{,}959$, con cuatro categorías distribuidas de forma casi uniforme entre condiciones (*Black*: 4/12, *East Asian*: 3/12, *White*: 3/12, *Latino*: 2/12). La variable que concentra mayor atención regulatoria resulta ser la más sensible al contexto; la no protegida, la más rígida.

Para dotar de garantía inferencial a estos patrones, complementamos la entropía con intervalos de confianza sobre las proporciones extremas. La ausencia de *Suit & Tie* no es un accidente muestral: con 0 selecciones sobre 6.000, el límite superior del intervalo de confianza al 95 % (regla de tres) se sitúa en el 0,05 %, de modo que la probabilidad real de que el modelo seleccione atuendo formal para el perpetrador es, con alta confianza, inferior a 1 de cada 2.000. De forma simétrica, la selección de género masculino en las 6.000 respuestas deja el intervalo en $[99{,}95\,\%, 100\,\%]$. Estas cotas convierten la invariancia observada en vestimenta y género de una regularidad descriptiva en una afirmación estadísticamente respaldada.

La variabilidad de la raza tampoco es atribuible al muestreo. En Gemma3 27B, cuyas respuestas son deterministas, condiciones que solo difieren en idioma o localización producen distribuciones raciales completamente disjuntas —por ejemplo, *Black* en el 100 % de la condición genérica en inglés frente a *White* en el 100 % de Pekín en inglés—. La asociación entre contexto y raza es en estos casos máxima ($\chi^2 = 1000$, gl $= 1$, $p < 10^{-100}$ para $n = 500$ por condición), lo que descarta que el desplazamiento de la categoría dominante sea ruido de generación.

5. Discusión

Los resultados presentados en la sección anterior permiten contrastar empíricamente la distinción conceptual introducida en la sección 2: la diferencia entre sesgo contextual y sesgo estructural.

5.1. Dos formas de sesgo: contextual y estructural

Los resultados ofrecen una ilustración empírica nítida de los dos patrones definidos en la sección 2. La raza exhibe con claridad el sesgo contextual: ante la misma escena de robo a mano armada, los modelos producen perfiles raciales distintos según el idioma del prompt y la ciudad en que se sitúa la escena. El sesgo existe, pero su contenido es negociable, cambia cuando cambia el anclaje cultural activado por el texto. La vestimenta, en cambio, ilustra el sesgo estructural de forma extrema: *Hoodie&Baggy Pants* domina en las 12 condiciones del diseño, sin que la categoría *Suit&Tie* aparezca en ninguna de las 6.000 respuestas.

5.2. Variables no protegidas como objetos independientes de preocupación

Los resultados empíricos permiten volver al marco de la sección 2. Si la práctica técnica tiende a hacer visibles las variables no protegidas principalmente como proxies de atributos protegidos, el caso de la vestimenta muestra por qué esa restricción es insuficiente. La asociación observada entre vestimenta y criminalidad no necesita operar como sustituto de raza o género para ser sistemática y socialmente relevante. Un sistema que genera descripciones de sospechosos o perfiles narrativos puede estar codificando, de forma consistente y sin activar mecanismos de alerta, asociaciones entre determinadas formas de vestir y conducta delictiva. El daño potencial no depende solo de si el atributo está o no protegido legalmente: depende de si la asociación es sistemática, de si se reproduce en aplicaciones reales y de si tiene capacidad de influir en decisiones o representaciones que afectan a personas. El mapa de atributos auditados por la práctica técnica puede ser, por tanto, demasiado estrecho para capturar algunas formas de estigmatización generadas por LLMs.

5.3. Implicaciones para la evaluación de sesgos en LLMs

Los hallazgos apuntan a tres reorientaciones. La primera es ampliar el espectro de variables analizadas más allá de los atributos protegidos clásicos: si las auditorías solo examinan raza, género, religión u orientación sexual, dejan fuera atributos sobre los que los modelos pueden generar asociaciones sistemáticas, como vestimenta, acento, lugar de procedencia o marcadores corporales. La segunda es incorporar métricas de estabilidad contextual junto a las métricas de disparidad entre grupos: un sistema de evaluación que solo mide si el modelo trata de forma diferente a dos grupos ante el mismo input está diseñado para detectar sesgo contextual, pero puede ser ciego al sesgo estructural, cuya detección requiere examinar si la distribución de respuestas permanece invariante ante variaciones sistemáticas del contexto. La tercera afecta a los mecanismos de mitigación: los esfuerzos de alineación actuales —como, por ejemplo, filtros de contenido y prompts de seguridad— se diseñan y evalúan principalmente sobre variables protegidas, y los resultados sugieren que esos mecanismos no garantizan por sí mismos la eliminación de asociaciones rígidas sobre variables no

protegidas. Los protocolos de alineación deben extenderse explícitamente a estas variables, y su efectividad debe evaluarse también sobre ellas.

5.4. Limitaciones

El estudio presenta varias limitaciones. Cubre un único tipo de tarea —perfil de sospechoso en contexto de casting— y los resultados no son directamente generalizables a otros dominios. La muestra de modelos es reducida, lo que limita la capacidad de extrapolar los patrones al conjunto de LLMs disponibles. El diseño incluye solo dos idiomas y tres localizaciones, una cobertura más amplia permitiría caracterizar con mayor precisión la estructura del sesgo contextual. El formato de respuesta cerrado en JSON acota además el espacio de salida a un conjunto predefinido de categorías, por lo que las proporciones observadas deben leerse como dominancias relativas dentro de ese menú y no como frecuencias absolutas en generación libre. Finalmente, el encuadre narrativo de casting puede introducir un sesgo de *framing*: la apelación a la verosimilitud narrativa podría activar asociaciones estereotípicas que no aparecerían en tareas más abstractas.

6. Conclusiones

Este trabajo ha examinado si las variables no protegidas pueden constituir objetos independientes de preocupación discriminatoria en LLMs, y si la naturaleza del sesgo que exhiben difiere cualitativamente del observado en variables protegidas. Los resultados confirman ambas hipótesis de forma contraintuitiva: la raza, variable que concentra mayor atención regulatoria, resulta ser la más sensible al contexto ($H = 1{,}959$), mientras que la vestimenta permanece invariante en las 12 condiciones del diseño ($H = 0{,}000$), sin que ninguna señal cultural, geográfica o lingüística logre perturbarla.

Este contraste sostiene la distinción entre sesgo contextual y sesgo estructural: el primero es modulable y detectable mediante diseños comparativos que manipulan atributos del prompt, mientras que el segundo opera por debajo del nivel en que esos instrumentos tienen capacidad de intervención y puede, por ello, resultar discriminatoriamente más grave en los contextos en que organiza representaciones socialmente perjudiciales.

La conclusión práctica es directa: los sistemas de auditoría, los marcos regulatorios y los protocolos de alineación deben ampliar el espectro de variables consideradas más allá de las categorías jurídicamente protegidas, e incorporar métricas orientadas a detectar invariancia y no solo disparidad. Como trabajo futuro, proponemos extender el diseño a más modelos, idiomas, localizaciones y tareas, así como analizar otras variables no protegidas, como acento, apariencia física, edad percibida, clase social aparente o nombres de pila.

Agradecimientos. Esta investigación es parte del Proyecto TSI-100927-2023-1, financiado por el Plan de Recuperación, Transformación y Resiliencia Europeo, de la Unión Europea-NextGenerationEU a través del Ministerio para la

Transformación Digital y de la Función Pública, y está parcialmente financiada por el proyecto PID2022-137953OB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE.

Declaración de intereses. Los autores declaran no tener conflictos de intereses que sean relevantes para el contenido de este artículo.

Bibliografía

[1] Bai, X., Wang, A., Sucholutsky, I., Griffiths, T.L.: Explicitly unbiased large language models still form biased associations. Proceedings of the National Academy of Sciences **122**(8), e2416228122 (2025)
[2] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 610–623 (2021)
[3] Chauhan, A., Anand, T., Jauhari, T., Shah, A., Singh, R., Rajaram, A., Vanga, R.: Identifying race and gender bias in stable diffusion ai image generation. In: 2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC). pp. 1–6. IEEE (2024)
[4] Ding, F., Hardt, M., Miller, J., Schmidt, L.: Retiring adult: New datasets for fair machine learning. Advances in neural information processing systems **34**, 6478–6490 (2021)
[5] Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K.: Bias and fairness in large language models: A survey. Computational linguistics **50**(3), 1097–1179 (2024)
[6] Gerards, J., Borgesius, F.Z.: Protected grounds and the system of nondiscrimination law in the context of algorithmic decision-making and artificial intelligence. Colo. Tech. LJ **20**, 1 (2022)
[7] Hartley, A., Fisher, J.: Seeing is believing? exploring gender bias in artificial intelligence imagery of specialty doctors. The Clinical Teacher **23**(1), e70297 (2026)
[8] Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1548–1558 (2021)
[9] Kotek, H., Dockum, R., Sun, D.: Gender bias and stereotypes in large language models. In: Proceedings of the ACM collective intelligence conference. pp. 12–24 (2023)
[10] Kreso, A., Boban, Z., Kabic, S., Rada, F., Batistic, D., Barun, I., Znaor, L., Kumric, M., Bozic, J., Vrdoljak, J.: Using large language models as decision support tools in emergency ophthalmology. International Journal of Medical Informatics **199**, 105886 (2025)
[11] Lo, F.P.W., Qiu, J., Wang, Z., Yu, H., Chen, Y., Zhang, G., Lo, B.: Ai hiring with llms: A context-aware and explainable multi-agent framework

for resume screening. In: Proceedings of the computer vision and pattern recognition conference. pp. 4184–4193 (2025)
[12] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
[13] Rabonato, R.T., Berton, L.: A systematic review of fairness in machine learning. AI and Ethics **5**(3), 1943–1954 (2025)
[14] Rudinger, R., Naradowsky, J., Leonard, B., Van Durme, B.: Gender bias in coreference resolution. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 8–14 (2018)
[15] Salinas, A., Haim, A., Nyarko, J.: What's in a name? auditing large language models for race and gender bias. arXiv preprint arXiv:2402.14875 (2024)
[16] Sasseville, M., Ouellet, S., Rhéaume, C., Sahlia, M., Couture, V., Després, P., Paquette, J.S., Darmon, D., Bergeron, F., Gagnon, M.P.: Bias mitigation in primary health care artificial intelligence models: scoping review. Journal of medical Internet research **27**, e60269 (2025)
[17] Tatman, R.: Gender and dialect bias in youtube's automatic captions. In: Proceedings of the first ACL workshop on ethics in natural language processing. pp. 53–59 (2017)
[18] Wachter, S.: The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law. Tul. L. Rev. **97**, 149 (2022)
[19] Wu, X., Wang, Q.R.: Popular llms amplify race and gender disparities in human mobility. arXiv preprint arXiv:2411.14469 (2024)
[20] Zeiser, J.: Emergent discrimination: Should we protect algorithmic groups? Journal of Applied Philosophy **42**(3), 910–928 (2025)
[21] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 conference on empirical methods in natural language processing. pp. 2979–2989 (2017)
[22] Zhao, T., Dai, E., Shu, K., Wang, S.: Towards fair classifiers without sensitive attributes: Exploring biases in related features. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. pp. 1433–1442 (2022)