Literacidad en IA: Funcionamiento, Sesgos y Operatividad de los LLMs

AI Literacy: How LLMs Work, Their Biases, and Their Operation

René Alexis Rodríguez Joaquín
https://orcid.org/0000-0002-0116-9196
renealexis.rodriguezjoaquin@gmail.com
Universidad de Guadalajara, México

DOI: https://doi.org/10.61454/yhetdk61

Resumen

Los modelos grandes de lenguaje (LLMs) han pasado de objeto de la lingüística computacional a infraestructura cotidiana de la producción profesional, científica y administrativa. Este ensayo propone una literacidad en IA entendida como comprensión técnico-epistémica del artefacto LLM, organizada en tres ejes: cómo funciona, qué sesgos inscribe y qué operatividad responsable cabe esperar de su uso. Desde un paradigma cualitativo-interpretativo, se adopta un diseño de investigación documental con enfoque crítico-analítico, siguiendo el modelo de Bowen (2009), a partir de literatura primaria indexada en Web of Science, Scopus, Google Scholar, arXiv y la ACL Anthology entre 2017 y mayo de 2026, sobre arquitecturas transformer (Vaswani et al., 2017), tokenización (Petrov et al., 2023; Ahia et al., 2023), aprendizaje por refuerzo con retroalimentación humana (Ouyang et al., 2022; Casper et al., 2023), validez de constructo (Raji et al., 2021; Bowman y Dahl, 2021) y colonialidad de datos (Quijano, 2000; Couldry y Mejias, 2019). El análisis identifica tres operaciones técnicas centrales (tokenización, atención y predicción del siguiente token) cuyas decisiones de diseño penalizan al español frente al inglés con un premium de tokenización de 1,55 a 1 en GPT-3.5 y GPT-4; tres capas estructurales de inscripción de sesgos (corpus, alineamiento por RLHF y benchmarks) que invalidan la pretensión de neutralidad de los sistemas comerciales; una caída de la transparencia agregada de los modelos de fundación de 58 a 40,69 puntos sobre 100 entre 2024 y 2025; y una brecha de hasta 30,2 puntos porcentuales entre la mejor y la peor lengua en MMLU-ProX. Se concluye que la literacidad técnico-epistémica es condición necesaria para que la academia, los organismos públicos, el sector privado y el tercer sector latinoamericanos articulen políticas de IA situadas, transparentes y soberanas.

Palabras Clave

literacidad en IA, modelos grandes de lenguaje, tokenización, sesgo algorítmico, RLHF, benchmarks, transparencia de datos.

Abstract

Large Language Models (LLMs) have moved from a specialized object of computational linguistics to a routine infrastructure for professional, scientific and administrative production. This essay proposes an AI literacy understood as the technical-epistemic comprehension of the LLM artifact, organized in three axes: how it works, what biases it inscribes, and what responsible operability can be expected from its use. From a qualitative-interpretive paradigm, a documentary research design with a critical-analytical approach is adopted, following Bowen's (2009) model, drawing on primary literature indexed

in Web of Science, Scopus, Google Scholar, arXiv and the ACL Anthology between 2017 and May 2026, on transformer architectures (Vaswani et al., 2017), tokenization (Petrov et al., 2023; Ahia et al., 2023), reinforcement learning from human feedback (Ouyang et al., 2022; Casper et al., 2023), construct validity (Raji et al., 2021; Bowman and Dahl, 2021), and data coloniality (Quijano, 2000; Couldry and Mejias, 2019). The analysis identifies three central technical operations (tokenization, attention and next-token prediction) whose design decisions penalize Spanish relative to English with a tokenization premium of 1.55 to 1 in GPT-3.5 and GPT-4; three structural layers of bias inscription (corpus, RLHF alignment and benchmarks) that invalidate the neutrality claim of commercial systems; a drop in aggregate transparency of foundation models from 58 to 40.69 points out of 100 between 2024 and 2025; and a gap of up to 30.2 percentage points between the best and worst languages in MMLU-ProX. The article concludes that technical-epistemic literacy is a necessary condition for Latin American academia, public bodies, private sector, and the third sector to articulate situated, transparent, and sovereign AI policies.

Keywords
AI literacy, large language models, tokenization, algorithmic bias, RLHF, benchmarks, corpus transparency.

Recepción: 6 de marzo de 2026
Aceptación: 19 de junio de 2026

Introducción

En menos de cuatro años, los modelos grandes de lenguaje han dejado de ser un objeto de investigación restringido a laboratorios especializados para convertirse en infraestructura de uso cotidiano en oficinas públicas, despachos jurídicos, centros de salud, redacciones periodísticas y aulas universitarias. Esta velocidad de adopción contrasta con el ritmo, más lento, al que se ha difundido el conocimiento sobre cómo funcionan estos sistemas y qué inscriben sus arquitecturas. El resultado es un desfase: muchas personas utilizan diariamente herramientas cuya descripción técnica básica está fuera de su repertorio conceptual, y muchas decisiones organizacionales sobre adopción de IA se toman a partir de demostraciones comerciales antes que de evidencia empírica.

La pregunta que motiva este trabajo es descriptiva antes que prescriptiva: ¿qué necesita saber, sobre el objeto LLM, una persona que pretende utilizarlo de forma profesional, científica o administrativa con responsabilidad? La hipótesis que se defiende es que tal conocimiento se articula en torno a tres ejes interdependientes que operan en orden lógico. Primero, el funcionamiento: las tres operaciones técnicas que producen una salida (tokenización, atención y predicción del siguiente token) y la genealogía de la arquitectura *transformer* que las hace posibles. Segundo, los sesgos: las tres capas estructurales donde el sistema deja de ser neutral (corpus, alineamiento y evaluación). Tercero, la operatividad: los criterios verificables con que un modelo concreto puede ser auditado, comparado o desestimado para un caso de uso particular.

La noción de literacidad que se propone diverge tanto de la versión funcionalista que la reduce a competencias de uso como de la versión generalista que la disuelve en una alfabetización digital genérica. La literacidad en IA, tal como aquí se entiende, es la capacidad de describir el artefacto LLM en términos técnicamente correctos sin perder rigor crítico sobre sus condiciones de producción. Requiere familiaridad con vocabulario de lingüística computacional, lectura de evidencia empírica

reciente y manejo de los marcos epistémicos que permiten evaluar la validez de las afirmaciones sobre IA. No equivale a programar un modelo ni a entrenarlo, pero sí a poder leer un reporte técnico, interpretar un *benchmark* y formular una pregunta auditable sobre un sistema desplegado.

La fundamentación teórica se ancla en cinco tradiciones convergentes. Primero, la lingüística computacional crítica de Bender y Koller (2020), que aporta la distinción entre forma lingüística y significado referencial. Segundo, los estudios decoloniales aplicados a la infraestructura de datos, en la estela de Quijano (2000), Mignolo (2002), Veronelli (2015), Ricaurte (2022) y Couldry y Mejias (2019). Tercero, la teoría de la medición psicométrica clásica (Cronbach y Meehl, 1955; Messick, 1995) extendida a la evaluación de la IA por Raji et al. (2021) y Bowman y Dahl (2021). Cuarto, los estudios de ciencia, tecnología y sociedad sobre infraestructuras digitales (Crawford, 2021; Birhane, 2020). Quinto, la filosofía de la información de Floridi (2013, 2023) y su lectura de los LLMs como agencia sin inteligencia.

La estructura del artículo refleja esta articulación. La segunda sección describe el enfoque metodológico de revisión documental crítica que sostiene la lectura. La tercera sección expone el funcionamiento del LLM desde la jerarquía conceptual de la inteligencia artificial, hasta el detalle de las tres operaciones centrales, con énfasis en la tokenización como sitio donde se inscribe la asimetría lingüística que paga el español. La cuarta sección examina los sesgos en tres capas: corpus de entrenamiento y opacidad documental, alineamiento por RLHF y benchmarks de evaluación. La quinta sección desarrolla la operatividad como práctica auditiva: cómo seleccionar un modelo, cómo verificar una respuesta, cómo medir el rendimiento por dominio. La sexta sección discute las implicaciones para América Latina. La séptima sección sintetiza las conclusiones y proyecta líneas futuras.

Metodología

El presente trabajo se inscribe en el paradigma cualitativo-interpretativo y adopta un diseño de investigación documental con enfoque crítico-analítico (Denzin y Lincoln, 2012). No constituye una revisión sistemática en sentido PRISMA, sino una revisión documental crítica orientada a la construcción argumentativa de un concepto, la literacidad en IA, a partir de literatura primaria interdisciplinar. El procedimiento de análisis documental sigue el modelo propuesto por Bowen (2009), articulado en cuatro fases: (i) localización y selección de fuentes, (ii) lectura analítica y codificación temática, (iii) síntesis interpretativa y (iv) contrastación crítica entre fuentes.

La búsqueda bibliográfica se realizó entre enero y abril de 2026 en cuatro fuentes complementarias: Web of Science y Scopus, como bases de datos multidisciplinares de alto factor de impacto; Google Scholar, como agregador de literatura gris y revistas latinoamericanas no indexadas en los anteriores; y arXiv junto con la ACL Anthology, como repositorios especializados que concentran la literatura primaria del campo del procesamiento del lenguaje natural y los modelos de fundación. Esta combinación responde a una característica del campo: la investigación técnica de referencia en IA generativa circula con frecuencia primero como preprint en arXiv o como actas de congreso (NeurIPS, ICML, ACL, NAACL, EMNLP, FAccT) antes de ser indexada en revistas tradicionales.

La estrategia de búsqueda se estructuró en torno a cinco bloques temáticos correspondientes a los ejes argumentales del artículo. El primero, arquitectura transformer y tokenización, combinó los descriptores “transformer architecture”, “tokenization fairness”, “subword tokenization” y “language model tokenizer”. El segundo, sesgos en corpus y alineamiento, combinó “RLHF bias”, “reinforcement

learning human feedback", "cultural bias LLM" y "WEIRD AI". El tercero, benchmarks y validez de constructo, combinó "benchmark validity", "MMLU contamination", "construct validity NLP" y "multilingual benchmark". El cuarto, transparencia y colonialidad de datos, combinó "data colonialism", "colonialidad de datos", "foundation model transparency" y "AI labor". El quinto, iniciativas latinoamericanas y literacidad en IA, combinó "LatamGPT", "Spanish language model", "literacidad en IA", "AI literacy" y "IA América Latina". Los descriptores se aplicaron en castellano e inglés con operadores booleanos (AND, OR) en cada motor de búsqueda.

El periodo temporal de la búsqueda contempló dos ventanas. Una ventana de literatura clásica, abierta hacia atrás sin restricción de fecha de inicio, destinada a recuperar los fundamentos teóricos de las cinco tradiciones convocadas: la psicometría de la validez de constructo (con Cronbach y Meehl, 1955, como referencia más antigua), la cibernética neuronal (Rosenblatt, 1958; Rumelhart, Hinton y Williams, 1986), los estudios decoloniales (Quijano, 2000; Mignolo, 2002) y la sociología de la comunicación (Goffman, 1981). Una ventana de literatura focal, comprendida entre 2017 (año de publicación de Attention Is All You Need, hito fundacional del paradigma transformer) y mayo de 2026, destinada a recoger la evidencia empírica reciente sobre arquitectura, tokenización, sesgos, benchmarks y transparencia. Esta segmentación permitió articular un anclaje conceptual estable con una lectura actualizada del estado del arte.

Los criterios de inclusión aplicados fueron cinco. Primero, literatura primaria publicada en revistas indexadas en Web of Science o Scopus, o en actas de congresos de referencia en el campo (ACL, NeurIPS, ICML, FAccT, NAACL, EMNLP). Segundo, reportes técnicos de modelos de fundación publicados por sus desarrolladores (OpenAI, Meta) o por observatorios independientes (Stanford Center for Research on Foundation Models). Tercero, literatura clásica de las tradiciones teóricas convocadas, aun cuando excediera la ventana focal. Cuarto, documentos publicados en español, inglés o portugués. Quinto, en el caso de fuentes periodísticas, únicamente cuando se trató de reportajes de investigación original con datos verificables que funcionaron como fuente primaria (por ejemplo, el reporte de *TIME* sobre las condiciones laborales en Sama-OpenAI).

Los criterios de exclusión eliminaron, en primer lugar, publicaciones de divulgación sin revisión por pares ni respaldo en datos verificables; en segundo lugar, entradas de blog corporativo de los proveedores de modelos cuando no estuvieron respaldadas por reportes técnicos paralelos; en tercer lugar, literatura secundaria que únicamente glosaba fuentes primarias ya incluidas en el corpus, en aplicación del principio metodológico de regreso a la fuente; y en cuarto lugar, materiales sin acceso al texto completo cuando la pieza no era localizable mediante repositorios institucionales ni vías bibliotecarias.

La organización de las fuentes se realizó mediante una matriz de codificación temática que registró, para cada referencia, cinco atributos: tradición teórica de procedencia, eje argumental al que aporta, tipo de evidencia (empírica, conceptual o técnica), región de procedencia institucional y año de publicación. Esta codificación permitió identificar zonas de saturación teórica y zonas de vacío, particularmente en lo referente a literatura latinoamericana sobre el objeto LLM. El procedimiento de análisis combinó tres operaciones articuladas: lectura analítica para identificar argumentos centrales y evidencia empírica de cada fuente; análisis de contenido cualitativo, en la línea de Krippendorff (2018), para extraer categorías recurrentes y construir las capas de inscripción de sesgos que articulan la sección homónima; y triangulación crítica entre fuentes técnicas, filosóficas y decoloniales para evitar el sesgo disciplinar y permitir la lectura situada que el artículo defiende.

Para garantizar la transparencia y la trazabilidad del procedimiento metodológico (Bowen, 2009), la organización de las fuentes se sistematizó mediante una matriz de codificación temática. A continuación, la Tabla 1 presenta un extracto representativo de este instrumento, ilustrando cómo se evaluaron los cinco atributos principales frente a fuentes de distinta naturaleza (literatura académica, teoría decolonial y reportes técnicos corporativos).

Tabla 1. *Extracto de la matriz de codificación temática para el análisis documental.*

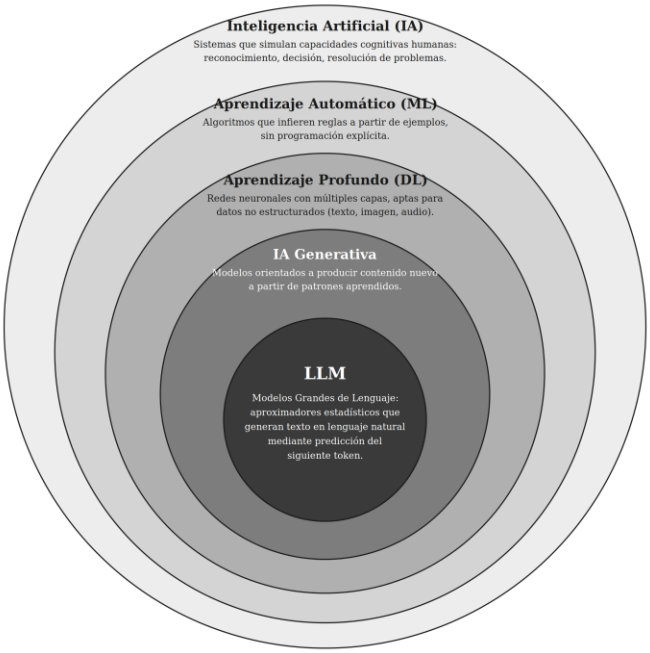
Referencia (Autor y Año)	Tradición teórica de procedencia	Eje argumental al que aporta	Tipo de evidencia	Región de procedencia institucional
Petrov et al. (2023)	Lingüística computacional	Funcionamiento (Asimetría en la tokenización del español)	Empírica / Técnica	Norte Global (Reino Unido / Arabia Saudita)
Bender y Koller (2020)	Lingüística computacional crítica	Funcionamiento (Distinción entre forma lingüística y significado)	Conceptual	Norte Global (Estados Unidos / Alemania)
Quijano (2000)	Estudios decoloniales latinoamericanos	Sesgos (Colonialidad del lenguaje y subordinación en corpus)	Teórica / Conceptual	América Latina (Perú)
Bommasani et al. (2025)	Evaluación métrica de IA (Stanford CRFM)	Sesgos e infraestructura (Opacidad en la trazabilidad del corpus)	Empírica	Norte Global (Estados Unidos)
OpenAI (2023) / Touvron et al. (2023)	Documentación corporativa / Reporte de la industria	Sesgos y operatividad (Estrategias de opacidad competitiva)	Técnica / Reporte comercial	Norte Global (Estados Unidos)

Funcionamiento de los LLMs: del perceptrón al transformer
Jerarquía conceptual: ubicar el objeto de estudio

Antes de discutir LLMs específicos conviene situarlos en una jerarquía conceptual cuya difuminación es fuente recurrente de errores divulgativos. La inteligencia artificial es el término más amplio y se refiere al diseño de sistemas que simulan capacidades cognitivas humanas como reconocimiento, decisión y resolución de problemas. El aprendizaje automático (*machine learning*) es un subconjunto de la IA que diseña algoritmos capaces de aprender de los datos sin ser programados explícitamente, infiriendo reglas a partir de ejemplos. El aprendizaje profundo (*deep learning*) es, a su vez, un subconjunto del aprendizaje automático que utiliza redes neuronales con múltiples capas, especialmente útil para datos no estructurados, como texto, imagen o audio, donde no es posible diseñar manualmente las características relevantes (Goodfellow et al., 2016).

Dentro del aprendizaje profundo se ubica la IA generativa, orientada a producir contenido nuevo a partir de patrones aprendidos. Y dentro de la IA generativa, los modelos grandes de lenguaje (LLMs) constituyen una familia particular orientada al texto, entrenada con decenas o cientos de miles de millones de parámetros sobre corpus de billones de tokens. Las herramientas comerciales, como ChatGPT, Claude, Gemini o DeepSeek, son las puntas visibles de un conjunto cuyo fondo es estadístico y matemático. Distinguir estos niveles permite no confundir una propiedad de la IA en general (por ejemplo, la capacidad de clasificar imágenes médicas) con una propiedad específica de los LLMs (la generación probabilística de texto). La Figura 1 sintetiza esta jerarquía.

Figura 1.
Jerarquía conceptual de la inteligencia artificial.



Cada nivel constituye un subconjunto del nivel inmediatamente superior. Adaptado de Goodfellow, Bengio y Courville (2016). Elaboración propia.

Genealogía: del perceptrón al transformer

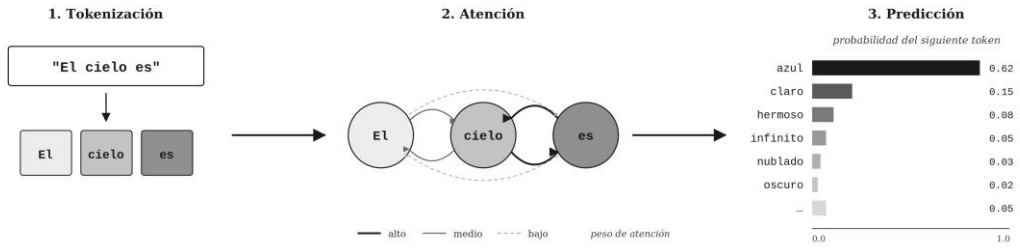
La historia técnica que conduce a los LLMs contemporáneos puede contarse en tres hitos: el perceptrón de Rosenblatt (1958) como antecedente de toda red neuronal; la retropropagación del error (*backpropagation*) de Rumelhart, Hinton y Williams (1986), que hace viable el aprendizaje profundo; y la arquitectura *transformer* de Vaswani et al. (2017), basada únicamente en mecanismos de atención. El *transformer* superó la limitación de las redes recurrentes al procesar toda la secuencia en paralelo y capturar dependencias a cualquier distancia, aunque con costo cuadrático en la longitud: esta restricción matemática es la que fija lo que hoy se conoce como ventana de contexto y explica por qué su dimensión se ha vuelto eje de competencia comercial (a mayo de 2026, Gemini y Claude superan el millón de tokens; Grok reporta hasta dos millones).

Tres operaciones centrales: tokenización, atención y predicción

Un LLM contemporáneo realiza tres pasos esenciales. Primero, el texto entrante se segmenta en tokens, unidades subléxicas producidas por algoritmos como BPE (Sennrich, Haddow y Birch, 2016), WordPiece (Devlin et al., 2019) o SentencePiece (Kudo y Richardson, 2018), y cada token se traduce a un vector denso de alta dimensión llamado *embedding* donde tokens semánticamente cercanos ocupan posiciones próximas (Sanderson, 2024). Segundo, la atención propia o *self-attention* pondera cada token según su relevancia respecto al resto del contexto, refinando su representación. Tercero, sobre esa representación contextualizada el modelo predice el siguiente token muestreando una distribución de probabilidad sobre el vocabulario, y el proceso se repite. El argumento técnico clave que sostiene cualquier descripción rigurosa del objeto es el siguiente: el *transformer* no entiende, redistribuye estadísticamente la información del contexto sobre cada token (Teixeira et al., 2025). Los pesos de atención son patrones de correlación distribucional aprendidos, no relaciones semánticas comprendidas. La Figura 2 sintetiza las tres operaciones en un flujo único.

Figura 2.

Las tres operaciones centrales del LLM: tokenización, atención y predicción del siguiente token.



El LLM no comprende ni razona en sentido referencial: redistribuye estadísticamente la información del contexto sobre cada token (Bender y Koller, 2020; Floridi, 2023). Las probabilidades mostradas son ilustrativas. Elaboración propia.

Tokenización: la asimetría que paga el español

Dentro de las tres operaciones, la tokenización merece tratamiento específico porque es donde se inscribe materialmente la primera asimetría política del sistema. En tokenizadores entrenados se sobrerrepresentan dominantemente los anglófonos, palabras inglesas frecuentes se codifican en un solo token, mientras que palabras en español, con mayor riqueza morfológica, tildes, ñ y desinencias, se fragmentan habitualmente. Petrov et al. (2023), en NeurIPS, documentaron esta asimetría sobre el corpus FLORES-200 y reportaron que el tokenizador de ChatGPT y GPT-4 (cl100k_base) emplea cerca de 1,55 veces más tokens para codificar el mismo texto en español que en inglés (ver tabla 2). Para el italiano la cifra es 1,6; para el búlgaro 2,6; para el árabe 3; y para el shan, lengua del estado birmano homónimo, hasta 15 veces más. Ahia et al. (2023) corroboran que los tokenizadores de las APIs comerciales favorecen desproporcionadamente las lenguas de escritura latina y fragmentan en exceso las lenguas y escrituras menos representadas.

Las implicaciones son tres y deben ser explicitadas en cualquier descripción rigurosa del objeto. Primera, los costos: como las APIs comerciales facturan por token consumido, una conversación en español cuesta alrededor de un 50 % más que la misma conversación en inglés. Segunda, el rendimiento: más fragmentación implica menos espacio semántico por palabra y, dado que la ventana de contexto se mide en tokens, su capacidad efectiva en palabras se reduce. Tercera, lo que importa

epistemológicamente: la subrepresentación está codificada en el tokenizador, antes incluso de que el modelo se invoque. No es una preferencia del modelo; es una decisión de diseño con consecuencias materiales para los hablantes de lenguas no anglófonas, especialmente las indígenas, donde la fragmentación es aún más severa.

Tabla 2. *Premium de tokenización del español frente al inglés en tokenizadores comerciales.*

Tokenizador	Modelos asociados	Premium español/inglés
r50k / p50k_base	GPT-2, RoBERTa	1,99×
cl100k_base	GPT-3.5, GPT-4	1,55×
FlanT5	Familia Flan-T5	2,23×

Nota. Adaptado de Petrov, La Malfa, Torr y Bibi (2023, Tabla 2, p. 3), evaluado sobre el corpus FLORES-200.

Forma versus significado: el debate filosófico de fondo

La distinción más útil para describir lo que un LLM hace es la que Bender y Koller (2020) formalizan en *Climbing towards NLU*: un sistema entrenado solo sobre forma no tiene, a priori, modo de aprender significado. Su *octopus test* ilustra el punto: un pulpo que intercepta cables submarinos entre dos islas y aprende a generar mensajes plausibles sin haber experimentado el mundo no comprende, y esto es estructuralmente análogo a lo que hace un LLM. La crítica converge con Bender et al. (2021), Mitchell y Krakauer (2023), Floridi (2023) y La Fontaine (2024): los LLMs son loros estocásticos que manipulan forma lingüística con sofisticación sin precedentes y producen efectos reales sobre los usuarios, pero atribuirles comprensión en sentido referencial es una categoría errada. Floridi (2023) lo sintetiza con precisión al describirlos como agencia sin inteligencia.

Sesgos: tres capas de inscripción

La pretensión de que los LLMs son herramientas neutrales es, a la luz de la literatura empírica reciente, insostenible. Los sesgos no son fallos accidentales del sistema; son consecuencias estructurales de cómo se construyen estos modelos en cada una de sus etapas. Esta sección identifica tres capas donde los sesgos se inscriben con efectos medibles: el corpus de entrenamiento, el alineamiento por RLHF, los *benchmarks* de evaluación.

Capa 1: el corpus y la opacidad estructural

Un corpus de pre-entrenamiento contemporáneo es un conjunto de varios billones de tokens compuesto típicamente por *web scraping* masivo, libros, código, literatura académica y foros. La fuente dominante es *Common Crawl*, cuyo archivo más reciente reporta cerca del 41 al 43 por ciento de contenido en inglés, frente a porcentajes notablemente menores para el español y el portugués. La cifra proxy más relevante para una audiencia latinoamericana proviene de los documentos técnicos de los propios modelos. Brown et al. (2020), en su artículo de presentación de GPT-3, declaran que la lengua del modelo es primariamente el inglés, con un 93 por ciento del recuento por palabra. En LLaMA 2, según el reporte técnico de Touvron et al. (2023), el inglés ocupa el 89,70 por ciento del corpus, mientras que el español apenas el 0,13 por ciento y el portugués el 0,09 por ciento.

Esta jerarquía no es un accidente técnico sino, como argumentan Quijano (2000) y Couldry y Mejias (2019), la consolidación computacional de la colonialidad del poder y de un nuevo colonialismo de datos. La taxonomía de Joshi et al. (2020) clasifica las cerca de siete mil lenguas vivas del planeta en

seis clases que van de *The Left-Behinds a The Winners*; el inglés y el español están en clase 5, pero el español es un ganador asimétrico, subordinado al inglés en los corpus y desproporcionadamente dominante respecto a las lenguas indígenas latinoamericanas, que casi en su totalidad pertenecen a la clase 0. Veronelli (2015), desde la filosofía decolonial latinoamericana, ofrece la categoría precisa: la colonialidad del lenguaje opera mediante un monolenguaje que naturaliza ciertos hablantes como sujetos legítimos de comunicación y vuelve invisibles a otros.

La transparencia sobre los corpus ha retrocedido. El *Foundation Model Transparency Index 2025* (Stanford CRFM) reporta una caída de la transparencia promedio a 40,69 puntos sobre 100, frente a 58 puntos en 2024 (Bommasani et al., 2025). Meta retrocedió de 60 a 31; Mistral, de 55 a 18. La invocación corporativa de un panorama competitivo, la fórmula que aparece en el reporte técnico de GPT-4 (OpenAI, 2023), funciona como un discurso legitimador que protege intereses corporativos del escrutinio público. Esta opacidad incluye, además del contenido, el trabajo invisible. La investigación de Perrigo (2023) en TIME documentó que OpenAI, mediante la subcontratista Sama, empleó trabajadores kenianos a entre 1,32 y 2 dólares por hora para etiquetar contenido tóxico durante el entrenamiento de los filtros de ChatGPT. Crawford (2021), en Atlas of AI, ofrece el marco para leer estas cadenas materiales: la IA es un registro de poder cuya fenomenología amable reposa en última instancia sobre minas, almacenes, centros de datos y mercados de trabajo precarios.

Capa 2: el RLHF y el espejismo de la neutralidad

El *Reinforcement Learning from Human Feedback* (aprendizaje por refuerzo a partir de retroalimentación humana, RLHF) es el procedimiento mediante el cual los modelos pre-entrenados son afinados para alinearse con preferencias humanas. Tres procesos interconectados componen esta cadena: recolección de *feedback*, modelado de recompensa y optimización de política (Casper et al., 2023). Cada etapa introduce sesgos. La pretensión corporativa de que el RLHF produce modelos neutrales o alineados con valores humanos universales es, a la luz de la evidencia, una afirmación ideológica.

Santurkar et al. (2023), en su trabajo sobre OpinionQA, cuantifican el sesgo de los modelos alineados como comparable a la división demócrata-republicana sobre el cambio climático. Durmus et al. (2023), en GlobalOpinionQA, muestran sesgo hacia Estados Unidos y algunos países europeos y sudamericanos. Atari et al. (2023) reportan una correlación de Pearson de −0,70 ($p < 0,001$) entre la distancia cultural respecto a Estados Unidos y la similitud de respuestas de GPT con poblaciones humanas. Los modelos alineados sobrerrepresentan a poblaciones WEIRD (occidentales, educadas, industrializadas, ricas y democráticas, en la formulación de Henrich, Heine y Norenzayan, 2010), especialmente estadounidenses de clase media urbana de orientación liberal. Cuando los datos de RLHF se recolectan casi exclusivamente en inglés, ese alineamiento no se transfiere: Havaldar et al. (2023) muestran que los LLMs multilingües siguen comportándose WEIRD e incluso al responder en lenguas no anglófonas.

La cadena laboral añade otra dimensión. Las instrucciones de anotación que reciben los trabajadores en Nairobi, Manila o Caracas son redactadas por equipos de producto en San Francisco. Lo que cuenta como tóxico, sesgado, útil o inapropiado se decide en el Norte y se aplica mecánicamente en el Sur. Cuando una instrucción dice marque como sesgada toda referencia a género asignada por defecto, el anotador no tiene canal para impugnar la guía explicando que en su contexto sociolingüístico el binario funciona distinto. La calidad de los datos así producidos, presentada como reflejo de valores humanos, es reflejo de la cosmovisión y las prioridades comerciales de quienes diseñan las directrices.

Capa 3: *benchmarks* y validez de constructo

La tercera capa de sesgo opera en la evaluación. Los *benchmarks* dominantes, como MMLU, HellaSwag, HumanEval, GSM8K o BBH, son las métricas con que se compara modelos en publicaciones académicas y en mensajes de marketing. La literatura crítica reciente muestra que estos *benchmarks* miden mucho menos de lo que dicen medir. Gema et al. (2024) estiman un 6,49 por ciento de errores en MMLU, con tasas de hasta 57 por ciento en virología; Balloccu et al. (2024) documentan que cerca de 4,7 millones de muestras de 263 *benchmarks* han estado expuestas a GPT-3.5 y GPT-4, lo que invalida la separación entre conjuntos de entrenamiento y evaluación; Pezeshkpour y Hruschka (2024) muestran caídas de 13 a 75 por ciento sólo reordenando las opciones de respuesta múltiple.

Desde la teoría de la medición clásica (Cronbach y Meehl, 1955; Messick, 1995), un instrumento mide bien un constructo cuando hay un constructo bien definido, una operacionalización fiel y ausencia de varianza irrelevante al constructo. Los *benchmarks* actuales fallan en los tres frentes. Raji et al. (2021), en su crítica fundacional, argumentan que los *benchmarks* generales operan como *stand-ins* que se asumen capturan capacidades generales, pero ningún *dataset* finito puede contener un constructo abstracto como inteligencia o entendimiento. Bowman y Dahl (2021) sostienen, en la misma línea, que la evaluación de la comprensión del lenguaje natural está rota.

Operatividad: la literacidad como práctica auditiva

Comprender el funcionamiento de un LLM y reconocer sus sesgos no agota la literacidad. Falta el tercer eje: saber qué hacer con esa comprensión cuando un sistema concreto debe ser elegido, comparado, desplegado o desestimado para un caso de uso particular. Este eje es el que aquí se denomina operatividad, no en el sentido de aprender a teclear instrucciones efectivas, sino en el sentido auditivo de poder formular preguntas verificables sobre el rendimiento, la trazabilidad y los costos reales de un modelo. La operatividad así entendida es una práctica de auditoría continua que el usuario informado realiza sobre el sistema con el que trabaja.

Cinco preguntas auditables

Una literacidad operativa puede ordenarse en torno a cinco preguntas que cualquier profesional, investigador o tomador de decisiones debería poder formular antes de adoptar un LLM en un flujo de trabajo crítico. Cada pregunta tiene una respuesta empíricamente verificable y un protocolo asociado.

La primera pregunta es por la trazabilidad del corpus: ¿sabe el proveedor qué datos entrenaron al modelo, en qué proporciones por idioma y de qué fuentes? La respuesta puede consultarse en el *Foundation Model Transparency Index* (Bommasani et al., 2025); cuando la respuesta es opaca, la decisión racional es asumir que cualquier reclamo de imparcialidad es indemostrable.

La segunda pregunta es por el rendimiento por idioma y dominio: ¿cuál es la precisión del modelo en español y en el dominio específico del caso de uso? Jackson et al. (2025), en su *Omniscience Index*, muestran que la diferencia entre la mejor y la peor opción en un dominio dado puede superar los cuarenta puntos: la práctica auditiva exige resultados desagregados, no agregados globales.

La tercera pregunta es por la calibración: ¿sabe el modelo cuándo no sabe? La métrica relevante no es la tasa de aciertos sino la de respuestas incorrectas presentadas con alta confianza. Los reportes de Artificial Analysis documentan que ciertos modelos comerciales mantienen entre el 39 y el 81 por ciento de respuestas inventadas en dominios técnicos; una literacidad operativa exige verificar muestras aleatorias y nunca delegar la verificación final al propio modelo que produjo la respuesta.

La cuarta pregunta es por los costos reales del español: el premium de tokenización documentado por Petrov et al. (2023) implica que organizaciones latinoamericanas pagan, a igualdad de funcionalidad,

cerca de un 50 por ciento más que las anglófonas. Esta asimetría debe entrar en cualquier análisis de costo total de propiedad.

La quinta pregunta es por la auditabilidad de la salida: ¿es posible reproducir la misma respuesta a partir del mismo prompt? En sistemas comerciales, generalmente no: la temperatura del muestreo y las actualizaciones silenciosas de los modelos *hosted* impiden la reproducibilidad estricta. Cualquier afirmación científica que cite resultados de LLMs exige documentar modelo, versión, parámetros de muestreo y fecha de consulta.

Implicaciones para América Latina

La literacidad en Inteligencia Artificial (IA) presenta desafíos específicos para América Latina, haciendo indispensable transitar de la dependencia corporativa hacia la soberanía tecnológica y lingüística.

Desafíos estructurales

- **Asimetría lingüística**: el español y el portugués ocupan roles subordinados a nivel global, mientras que cientos de lenguas indígenas continentales son invisibles en los corpus comerciales.
- **Dependencia tecnológica**: operamos bajo infraestructuras, costos y reglas dictadas por el Norte Global. Aunque existen iniciativas regionales valiosas (como LatamGPT, MarIA o Sabiá), muchas son solo adaptaciones (*fine-tunes*) de arquitecturas extranjeras. La soberanía efectiva requiere infraestructura propia, gobernanza local y datos consentidos.

El rol de la universidad latinoamericana

Las instituciones académicas tienen una triple responsabilidad para contextualizar el desarrollo de la IA:

1. **Formación**: educar profesionales con visión técnico-crítica, capaces de auditar la IA en sectores públicos y privados.
2. **Investigación**: generar conocimiento situado apoyándose en redes interinstitucionales consolidadas (Khipu, Latinx in AI, entre otras).
3. **Regulación**: colaborar con diferentes sectores para definir políticas propias de transparencia, auditoría y protección.

Líneas de acción y adopción

Para materializar estos objetivos, la región debe enfocarse en:

- Desarrollar capacidades de evaluación (*benchmarks* locales) que contemplen tanto las variedades regionales del español y portugués como las principales lenguas originarias.
- Priorizar modelos abiertos (*open-weights*) frente a APIs comerciales cerradas. Estos permiten realizar auditorías directas, proteger datos sensibles en servidores locales y reducir la dependencia de terceros.
- Fomentar la cooperación regional a través de marcos como el plan UNESCO-CEPAL. Ningún país debe asumir solo los altos costos de infraestructura y desarrollo.

El éxito regional no depende de decidir si adoptar o no la IA, sino de cómo adoptarla. Esto exige una articulación operativa real entre la academia, el gobierno y la sociedad civil, guiada por una literacidad técnico-epistémica basada en evidencias compartidas.

Conclusiones

Este artículo ha sostenido que la literacidad en IA, entendida como comprensión técnico-epistémica del artefacto LLM, constituye una competencia académica, profesional y ciudadana de primer orden. Tres conclusiones articulan el argumento desarrollado.

Primera, en el plano del funcionamiento: los LLMs no son razonadores ni comprendedores en sentido referencial sino aproximadores estadísticos compuestos por tres operaciones técnicas (tokenización, atención y predicción del siguiente token) cuyas decisiones de diseño tienen consecuencias materiales medibles. La asimetría de tokenización penaliza al español frente al inglés en una proporción de 1,55 a 1 (Petrov et al., 2023), y la distinción entre forma y significado fijada por Bender y Koller (2020) sostiene cualquier descripción rigurosa del objeto.

Segunda, en el plano de los sesgos: la pretensión de neutralidad de los modelos comerciales se desmiente empíricamente en tres capas. El corpus es opaco y dominantemente angloparlante, y la transparencia agregada cayó del 58 al 40,69 sobre 100 entre 2024 y 2025 (Bommasani et al., 2025); el RLHF inyecta preferencias WEIRD bajo apariencia universal (Santurkar et al., 2023; Atari et al., 2023); los *benchmarks* miden constructos pobremente definidos en condiciones culturalmente sesgadas, con brechas entre lenguas de hasta 30,2 puntos (Xuan et al., 2025).

Tercera, en el plano de la operatividad: una literacidad madura no se reduce al uso instrumental, sino que se despliega como práctica auditiva mediante cinco preguntas verificables (trazabilidad del corpus, rendimiento por idioma y dominio, calibración, costos reales del español y auditabilidad de la salida) y se distribuye en tres niveles de profundidad práctica (usuario informado, equipo organizacional y actor regulatorio) que articulan responsabilidades crecientes.

Las implicaciones para América Latina son significativas. La literacidad en IA contribuye al desarrollo profesional formando ciudadanos capaces de operar con autonomía crítica en entornos digitales mediados por IA; al avance de la ciencia, generando investigación situada que produzca conocimiento original sobre IA en español y portugués; y a la transformación social, creando condiciones para que las herramientas de IA fortalezcan, y no erosionen, la soberanía cognitiva regional. Los puentes con organismos públicos, privados y del tercer sector encuentran aquí un objeto compartido: una IA situada, justa, transparente y útil.

La adopción de la IA generativa avanza más rápido que su descripción rigurosa. Comprender el LLM no es opcional: es la condición mínima para que el uso masivo no se confunda con el saber. Si la academia latinoamericana asume esta tarea descriptiva con el rigor que merece, podrá aportar al campo internacional algo más que adopción tardía: podrá aportar evidencia situada, perspectiva crítica y propuestas regulatorias propias. La literacidad en IA, así entendida, no es un suplemento al currículum sino una condición de posibilidad de la investigación responsable y de la ciudadanía en la era algorítmica.

Referencias

Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D. R., Smith, N. A., y Tsvetkov, Y. (2023). Do all languages cost the same? Tokenization in the era of commercial language models. En *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9904-9923). Association for Computational Linguistics.

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., y Henrich, J. (2023). *Which humans?* PsyArXiv. https://doi.org/10.31234/osf.io/5b26t

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022). *Constitutional AI: Harmlessness from AI feedback*. arXiv. https://doi.org/10.48550/arXiv.2212.08073

Balloccu, S., Schmidtová, P., Lango, M., y Dušek, O. (2024). Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. En *Proceedings of the 18th Conference of the European Chapter of the ACL* (Vol. 1, pp. 67-93). Association for Computational Linguistics.

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

Bender, E. M., y Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. En *Proceedings of the 58th Annual Meeting of the ACL* (pp. 5185-5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed*, *17*(2), 389-409. https://doi.org/10.2966/scrip.170220.389

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Liang, P., et al. (2025). *Foundation Model Transparency Index 2025*. arXiv. https://doi.org/10.48550/arXiv.2512.10169

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27-40. https://doi.org/10.3316/QRJ0902027

Bowman, S. R., y Dahl, G. E. (2021). What will it take to fix benchmarking in natural language understanding? En *Proceedings of the 2021 Conference of the North American Chapter of the ACL: Human Language Technologies* (pp. 4843-4855). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.385

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. En *Advances in Neural Information Processing Systems*, 33 (pp. 1877-1901). Curran Associates.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., et al. (2023). *Open problems and fundamental limitations of Reinforcement Learning from Human Feedback*. arXiv. https://doi.org/10.48550/arXiv.2307.15217

Couldry, N., y Mejias, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television and New Media*, *20*(4), 336-349. https://doi.org/10.1177/1527476418796632

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Cronbach, L. J., y Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302. https://doi.org/10.1037/h0040957

Denzin, N. K., y Lincoln, Y. S. (Coords.). (2012). *Manual de investigación cualitativa: Vol. IV. Métodos de recolección y análisis de datos*. Editorial Gedisa.

Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. En *Proceedings of the 2019 Conference of the NAACL: HLT* (Vol. 1, pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Durmus, E., Nyugen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., et al. (2023). *Towards measuring the representation of subjective global opinions in language models*. arXiv. https://doi.org/10.48550/arXiv.2306.16388

Floridi, L. (2013). *The ethics of information*. Oxford University Press.

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy and Technology*, *36*(15). https://doi.org/10.1007/s13347-023-00621-y

Gema, A. P., Leang, J. O. J., Hong, G., Devoto, A., Mancino, A. C. M., Saxena, R., et al. (2024). *Are we done with MMLU?* arXiv. https://doi.org/10.48550/arXiv.2406.04127

Goffman, E. (1981). *Forms of talk*. University of Pennsylvania Press.

Goodfellow, I., Bengio, Y., y Courville, A. (2016). *Deep learning*. MIT Press.

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575-599. https://doi.org/10.2307/3178066

Havaldar, S., Pressimone, S., Wong, E., y Ungar, L. H. (2023). Multilingual language models are not multicultural: A case study in emotion. En *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 202-214). Association for Computational Linguistics.

Henrich, J., Heine, S. J., y Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61-83. https://doi.org/10.1017/S0140525X0999152X

Jackson, R., Liu, Y., y Patel, A. (2025). Omniscience Index: Measuring calibrated knowledge in large language models. *Artificial Analysis Reports*.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., y Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. En *Proceedings of the 58th Annual Meeting of the*

ACL (pp. 6282-6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications.

Kudo, T., y Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. En *Proceedings of the 2018 Conference on EMNLP: System Demonstrations* (pp. 66-71). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-2012

La Fontaine, G. (2024). Sobre loros estocásticos: Una mirada a los modelos grandes de lenguaje. *Lógoi: Revista de Filosofía*, *45*, 75-87.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Mignolo, W. D. (2002). The geopolitics of knowledge and the colonial difference. *South Atlantic Quarterly*, *101*(1), 57-96. https://doi.org/10.1215/00382876-101-1-57

Mitchell, M., y Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

OpenAI. (2023). *GPT-4 technical report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. En *Advances in Neural Information Processing Systems*, 35 (pp. 27730-27744). Curran Associates.

Perrigo, B. (2023, 18 de enero). Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. *TIME*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Petrov, A., La Malfa, E., Torr, P. H. S., y Bibi, A. (2023). Language model tokenizers introduce unfairness between languages. En *Advances in Neural Information Processing Systems*, 36 (pp. 36963-36990). Curran Associates.

Pezeshkpour, P., y Hruschka, E. (2024). Large language models sensitivity to the order of options in multiple-choice questions. En *Findings of the ACL: NAACL 2024* (pp. 2006-2017). Association for Computational Linguistics.

Quijano, A. (2000). Colonialidad del poder, eurocentrismo y América Latina. En E. Lander (Comp.), *La colonialidad del saber: eurocentrismo y ciencias sociales. Perspectivas latinoamericanas* (pp. 201-246). CLACSO.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., y Hanna, A. (2021). AI and the everything in the whole wide world benchmark. En *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (Vol. 1).

Ricaurte, P. (2022). Ethics for the majority world: AI and the question of violence at scale. *Media, Culture and Society*, *44*(4), 726-745. https://doi.org/10.1177/01634437221099612

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386-408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., y Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536. https://doi.org/10.1038/323533a0

Sanderson, G. (2024). Attention in transformers, visually explained. 3Blue1Brown. https://www.3blue1brown.com/lessons/attention

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., y Hashimoto, T. (2023). Whose opinions do language models reflect? En *Proceedings of the 40th International Conference on Machine Learning* (pp. 29971-30004). PMLR.

Sennrich, R., Haddow, B., y Birch, A. (2016). Neural machine translation of rare words with subword units. En *Proceedings of the 54th Annual Meeting of the ACL* (Vol. 1, pp. 1715-1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngoc Khai, J., Vila-Suero, D., et al. (2025). Global-MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. En *Proceedings of the 63rd Annual Meeting of the ACL* (Vol. 1). Association for Computational Linguistics.

Teixeira, F., Almeida, R., y Silva, M. (2025). Tokenization, context windows, and the architecture of language models for educators. *Computers and Education: Artificial Intelligence*, *8*, 100245.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). *Llama 2: Open foundation and fine-tuned chat models*. arXiv. https://doi.org/10.48550/arXiv.2307.09288

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., y Polosukhin, I. (2017). Attention is all you need. En *Advances in Neural Information Processing Systems*, 30 (pp. 5998-6008). Curran Associates.

Veronelli, G. (2015). Sobre la colonialidad del lenguaje. *Universitas Humanística*, *81*, 33-58. https://doi.org/10.11144/Javeriana.uh81.scl

Xuan, W., Yang, K., Tan, Y., Tian, X., Qin, Y., Zhang, M., et al. (2025). *MMLU-ProX: A multilingual benchmark for advanced large language model evaluation*. arXiv. https://doi.org/10.48550/arXiv.2503.10497