SOBERANÍA TECNOLÓGICA Y TOKENIZACIÓN: EL SESGO DEL NORTE GLOBAL EN LOS LLMs MODERNOS

Daniel Octavio Ramírez Gómez

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2026

Cita	(Ramírez Gómez, 2026)
Referencia **Estilo APA 7 (2020)**	Ramírez Gómez, D., (2026). *Soberanía tecnológica y tokenización: el sesgo del norte global en los LLMs modernos.* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte X.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Tabla de contenido

Resumen

El procesamiento de información a escala global se ha transformado gracias a los Modelos de Lenguaje Extensos (LLMs), estos se establecen desde la arquitectura Transformer y sus mecanismos de atención. Empero, existe poca neutralidad dentro de este avance puesto que los datos de entrenamiento de estos modelos reflejan una predominancia de las perspectivas del Norte Global. Este aspecto genera un borramiento digital de las culturas, lenguas y saberes de regiones como América Latina. Este informe, analiza las posibilidades de acción frente a dicho sesgo a partir del curso *Google DeepMind: AI Research Foundations*, articulando sus contenidos técnicos con un marco teórico crítico en torno a tres categorías: los LLMs, la tokenización y la soberanía tecnológica. Metodológicamente, se adopta un enfoque cualitativo, descriptivo-analítico, desarrollado en tres fases: análisis sistemático del curso, revisión documental especializada y triangulación crítica módulo por módulo. Aquí se evidencia que el sesgo algorítmico opera desde las capas más fundamentales del entrenamiento —tokenización, embeddings, backpropagation— y que técnicas como el ajuste fino supervisado (SFT), LoRA y el aprendizaje por refuerzo con retroalimentación humana (RLHF) situado representan vías viables hacia una soberanía tecnológica contextualizada. De esta manera, se reflexiona ante la posibilidad de cerrar la brecha entre los LLMs modernos y las comunidades por fuera del norte global requiere gobernanza democrática, curaduría ética de datos y participación colectiva en el diseño de los modelos.

Palabras clave: modelos de lenguaje extensos, tokenización, sesgo algorítmico, soberanía tecnológica, Norte Global, arquitectura Transformer, diversidad cultural, inteligencia artificial.

Abstract

The processing of information at a global scale has been transformed by Large Language Models (LLMs), which are grounded in the Transformer architecture and its attention mechanisms. However, this advancement carries little neutrality, as the training data of these models predominantly reflects the perspectives of the Global North — a condition that produces a digital erasure of the cultures, languages, and knowledge systems of regions such as Latin America. This report analyzes possible courses of action to address this bias through the *Google DeepMind: AI Research Foundations* course, articulating its technical content within a critical theoretical framework built around three categories: LLMs, tokenization, and technological sovereignty. Methodologically, a qualitative, descriptive-analytical approach is adopted, developed in three phases: systematic analysis of the course, specialized documentary review, and critical module-by-module triangulation. The findings reveal that algorithmic bias operates from the most fundamental layers of model training — tokenization, embeddings, and backpropagation— and that techniques such as Supervised Fine-Tuning (SFT), LoRA, and situated Reinforcement Learning from Human Feedback (RLHF) represent viable pathways toward a contextualized technological sovereignty. In this way, the study reflects the possibility of closing the gap between modern LLMs and communities beyond the Global North, a process that demands democratic governance, ethical data curation, and collective participation in model design.

Keywords: large language models, tokenization, algorithmic bias, technological sovereignty, Global North, Transformer architecture, cultural diversity, artificial intelligence.

SOBERANÍA TECNOLÓGICA Y TOKENIZACIÓN: EL SESGO DEL NORTE GLOBAL EN LOS LLMs MODERNOS

Introducción

Los Modelos de Lenguaje Extensos (LLMs) marcaron un hito al permitir que se procesara información de manera masiva a partir de mecanismos de atención, esto cambia la mirada porque dejan de usarse de manera frecuente los RNNs[1] y los LSTMs[2] que procesaban palabras una por una, demorando los procesamientos de información. Así pues, dicho logro fue fundamentado en la arquitectura Transformer propuesta por Vaswani et al. (2017) donde los autores proponen que dicho modelo puede prescindir de la recurrencia y basarse en mecanismos de atención para establecer dependencias globales entre entrada y salida. Es decir, se propone que el modelo transformer tendrá en cuenta tanto las entradas como las salidas para comprender el contexto que se aloja en la red neuronal. Es importante tener en cuenta que los mecanismos de autoatención (self-attention) permiten procesar todas las palabras en paralelo, calcular la relación matemática y el peso de cada palabra dentro de las redes neuronales.

Esto, permite tener una visión más amplia frente a los procesamientos de información masiva. Sin embargo, dentro de dicho avance técnico, existe una falta de representatividad donde los datos de entrenamiento reflejan únicamente a quienes poseen los recursos para participar en la conversación digital (Bender et al., 2021). En este caso, se puede evidenciar que las dinámicas de poder alrededor del manejo de la información se encuentran en el norte global. Aspecto que no es ajeno ante el procesamiento de información a partir de la arquitectura transformer y el auge de los LLMs. Tal y como lo plantean los autores:

> ...los grandes conjuntos de datos basados en textos de Internet

[1] Red Neuronal Recurrente.
[2] Versión mejorada de las redes neuronales recurrentes con memoria a corto y largo plazo.

> sobrerrepresentan los puntos de vista hegemónicos y codifican sesgos que pueden resultar perjudiciales para las poblaciones marginadas. Al recopilar conjuntos de datos cada vez más grandes, corremos el riesgo de incurrir en una «deuda de documentación (Bender et al., 2021, p.1)

Dicha situación responde a factores hegemónicos, pero también a la anatomía del transformer, ya que estos no procesan palabras secuencialmente, sino que requieren una manera de saber el orden de las palabras que solo se logra a través de codificaciones posicionales. Las cuales, se ejecutan originalmente a partir de funciones matemáticas fijas que proporcionan información de posición absoluta y relativa a los vectores de palabras. En términos de lo que plantean Abeliuk y Gutiérrez (2021):

> El procesamiento de lenguaje natural juega un papel vital en muchos sistemas, desde el análisis de curriculums para la contratación, hasta asistentes virtuales y detección de spam. Sin embargo, el desarrollo y la implementación de la tecnología de NLP no es tan equitativo como parece. Aunque se hablan más de 7000 idiomas en todo el mundo, la gran mayoría de los avances tecnológicos son aplicados al inglés p. 19.

En este sentido y teniendo en cuenta un contexto donde la cultura del norte global no predomina, sino que convergen una serie de identidades y culturas diversas que, en la poscolonialidad, se consolidan desde lo ancestral y lo occidental; es importante repensar los sistemas neuronales que privilegian visiones hegemónicas del mundo y que desplazan los saberes de cada cultura desplazándolas hacia la periferia del desarrollo tecnológico (Henao & Hernández, 2023). En contextos latinoamericanos, la diversidad cultural, étnica, lingüística y de saberes alude a lo planteado anteriormente. No existe un país que sea exactamente igual al otro puesto que cada identidad se configura de manera diferente. En un

territorio como el colombiano, los sistemas neuronales se traducen en una invisibilización de las particularidades culturales y lenguas nativas porque estas no son sistemas neutrales.

Para comprender dicho contexto, es importante situarse a partir del concepto de capital cultural planteado por Pierre Bourdieu (1997), donde el autor propone un modelo de capital cultural incorporado que se evidencia en las formas de hablar, actuar y los esquemas de pensamiento. En este sentido, es necesario analizar los datos de entrenamiento de los LLMs que parten –en su gran mayoría– del norte global, convirtiendo este escenario en el capital cultural legítimo. Este aspecto es relevante para el análisis aquí planteado porque en América Latina las formas de expresión a través del lenguaje pueden variar de acuerdo con el país e incluso región en la que se encuentren y son tomadas como errores por los modelos de arquitectura transformer ya que no son parte del centro hegemónico. Lo cual, implica una barrera técnica y simbólica que no representa a quienes poseen un capital cultural distinto.

Teniendo en cuenta lo anterior, se identifica la necesidad de crear modelos con mayor alcance de interpretabilidad para generar mayor visibilidad en las particularidades culturales y lenguas nativas en los sistemas de IA, entendiendo esta como una matriz de atención que funciona como un mapa de calor en el que se puede identificar qué palabras está reconociendo el modelo para entender una frase. Si estos arquetipos son más interpretables, el arquitecto de datos tiene la responsabilidad de auditar esas distribuciones de atención para identificar dónde el algoritmo está ignorando o malinterpretando las variantes lingüísticas locales. Dichas distribuciones de atención se entienden como el peso numérico que se concentra en las palabras y que –en la cotidianidad con los modelos comunes- no procesa siempre el parlache (la jerga de una ciudad) o las lenguas nativas, por lo que termina dando respuestas genéricas.

En este sentido, este trabajo tiene como propósito analizar las diferentes posibilidades de acción que se pueden presentar en el trabajo con LLMs para reducir el sesgo en la IA. Esto, a partir del curso de Google DeepMind, analizando con cada módulo el caso propuesto, las posibilidades aplicables y las que no. Así, en primer lugar, se hace una breve descripción de la problemática y los objetivos del proyecto. Posterior a ello, se da paso a las categorías

esenciales para abordar esta temática y sus perspectivas teóricas. Finalmente, se da un desarrollo del proyecto basado en los módulos del curso mencionado anteriormente y un análisis del caso planteado alrededor de este.

Planteamiento del problema

Como se ha mencionado en el apartado anterior, existe un sesgo en el cual hay una subrepresentación de las culturas foráneas al norte global. Esto, debido a la tendencia que ese tiene en los grandes modelos de lenguaje al homogeneizar el conocimiento a partir de un entrenamiento que sólo tiene en cuenta ciertos sectores hegemónicos. De acuerdo con Bender et Al. (2021), un LLM no comprende el significado ni el contexto cultural, toma probabilidades estadísticas y combina las palabras; este fenómeno se conoce, según las autoras, como el "loro estocástico", ya que no existe una mente o intención real, solo es un ejercicio repetitivo y de amplificación de patrones sin distinciones éticas.

Lo anterior se traduce en una centralización de la información y una invisibilización de las diversas formas de expresión que existen en todo el mundo. Dicho fenómeno es conocido como borramiento digital, donde los LLMs son entrenados de manera masiva con información encontrada en la web que es habitada hegemónicamente por el norte global que tiene acceso a la tecnología, generando un sesgo donde un solo lenguaje absorbe todo el contenido que se exporta. De esta manera, el problema se traslada a la invisibilidad que tienen identidades, culturas, jergas y perspectivas de comunidades con un lenguaje minoritario o que provienen de las periferias. Tal y como lo plantean Bender et. Al (2021): " En el ámbito de la minería de datos de texto, los puntos de vista dominantes suelen estar sobrerrepresentados, lo que margina aún más las perspectivas de las minorías. [...] El lenguaje de quienes se encuentran marginados queda, en la práctica, borrado de los datos de entrenamiento" Pág. 613. Esto representa un riesgo en los LLMs puesto que -al no tener una verdadera comprensión del capital cultural- se puede replicar y amplificar fácilmente el sesgo.

Ahora bien, la problemática planteada aquí se ha identificado a partir del desarrollo del curso de Google DeepMind, donde se especifica claramente en los laboratorios cómo los LLMs son herramientas que contienen un sesgo proveniente desde su entrenamiento. En este

sentido, se justifica la necesidad de analizar como profesional las formas de entrenamiento programadas para estos grandes modelos de lenguaje, entendiéndolos como sistemas socioculturales. Además, es importante resaltar la importancia de hacer una curaduría y documentación de los datos por parte de quien asume el entrenamiento del LLM, tomando decisiones que sean incluyentes y culturalmente diversas. Este es un problema que se ha rescatado desde diferentes estudios, en este caso Bender et Al. (2021) hacen una solicitud de "… reorientación de las prioridades de investigación hacia arquitecturas y metodologías que fomenten una comprensión integral de la composición de los conjuntos de datos y las prácticas de curación, así como una delimitación más clara de las capacidades y limitaciones de los modelos de lenguaje" pág. 618. De esta manera, se da paso al análisis a partir de la pregunta de investigación: *¿es posible cerrar la brecha entre los LLMs modernos y las comunidades teniendo en cuenta el capital cultural del territorio colombiano?*

Objetivos

Objetivo general

Analizar diferentes posibilidades de acción en el trabajo con LLMs para reducir el sesgo en la IA a partir del curso de Google DeepMind.

Objetivos específicos

1. Identificar la disparidad de representación y los grupos culturales con mayores índices de sesgo en los modelos comerciales de IA.
2. Analizar, desde una perspectiva técnica, los factores que causan la subrepresentación de las culturas del Sur Global, examinando procesos de tokenización, la arquitectura de los Transformers, la configuración de embeddings y el entrenamiento de redes neuronales.
3. Reflexionar frente a posibles estrategias desde la ciencia de datos para reducir el sesgo en IA e incrementar su enriquecimiento cultural.

Metodología

Este proyecto parte desde un abordaje metodológico con enfoque cualitativo y se fundamenta a partir de un enfoque descriptivo y analítico con perspectiva crítica. Dicho proceso se desarrolla en tres fases que se conectan entre sí de la siguiente manera:

Fase 1

Aquí, se toma como unidad de análisis central el programa de formación de google skills: *Google DeepMind: AI Research Foundations.* Este curso se ha desarrollado y examinado de manera sistemática desde sus módulos, laboratorios y casos de estudio. A partir de ello, se identificaron problemas conceptuales que fueron un punto clave para esta investigación; la persistencia de los sesgos epistémicos y técnicos en el entrenamiento de redes neuronales que se enfocan principalmente en el Norte Global y que dejan de lado las culturas e identidades que se encuentran en regiones como América Latina y Colombia. En este sentido y teniendo en cuenta el hallazgo dentro del curso, la práctica y el contenido de cada módulo se someten a un análisis de contenido crítico para evaluar las formas de aplicabilidad y las limitaciones que estas pueden tener frente al caso de estudio.

Fase 2

Con el fin de robustecer este informe y evitar la autorreferencialidad del curso, se considera importante acudir a las técnicas de recolección de información. Para este trabajo, se utiliza la revisión documental y bibliográfica especializada, buscando en bases de datos, repositorios, libros y revistas académicas. A partir de ello, se realizan fichas bibliográficas para la sistematización de la información y se establece un corpus teórico codificado en toro a categorías como los LLM, la tokenización, la soberanía tecnológica, las arquitecturas transformer y el sesgo del entrenamiento de IA. Esta selección de textos brinda la posibilidad de triangular la información entre los contenidos de google skills, los trabajos académicos especializados y el ejemplo del caso propuesto.

Fase 3

El desarrollo final del informe se estructura a partir de un desglose descriptivo-analítico de los módulos de Google skills. Aspecto que permite contrastar, módulo por módulo, las acciones técnicas del curso ante el componente teórico donde se discuten las posibilidades de aplicabilidad, las diferencias conceptuales con la producción académica sobre el tema y las adaptaciones necesarias para responder a la problemática del sesgo algorítmico en contextos locales.

Marco teórico

Este apartado busca dar una mirada al enfoque conceptual que, dentro del análisis del proceso formativo de Google skills, se considera más pertinente para comprender -desde lo epistémico- que el desarrollo de la IA no solo es un reto técnico y de entrenamiento de redes neuronales. Dicho aspecto, implica cuestionar también el carácter ético y geopolítico, trascendiendo la descripción instrumental del abordaje de los módulos. De esta manera, se estructura un marco teórico que aborda la dimensión técnica y la perspectiva crítica; desglosada en categorías de análisis como los LLMs, la Tokenización y la Soberanía Tecnológica. En efecto, estos conceptos no funcionan de forma aislada, son herramientas interconectadas para la comprensión y la evaluación de las posibilidades de acción y los límites del entrenamiento de redes neuronales.

Modelos de lenguaje de gran tamaño: LLMs

Este concepto ha sido estudiado y definido desde diferentes ámbitos de la Inteligencia Artificial ya que funcionan desde el entrenamiento, la predicción y la estructura, convirtiéndose así en un tipo de IA de aprendizaje profundo. Para delimitar este concepto fue necesario acceder a diferentes fuentes. Primero, se encuentra la percepción de Rodríguez et al. (2024) en su trabajo *Inteligencia artificial en la educación: Modelo de lenguaje de gran tamaño (LLM) como recurso educativo,* donde los autores definen los LLMs como sistemas avanzados de IA que pueden comprender y generar el lenguaje humano a partir del

procesamiento de grandes cantidades de información. Aquí, se estabece que los modelos tienen técnicas de aprendizaje profundo para mejorar las formas en las que se procesa el lenguaje natural. Esta percepción sobre los grandes modelos de lenguaje no dialoga con los objetivos categóricos de este trabajo porque su definición no diversifica las fuentes de información que son utilizadas para el entrenamiento de la IA.

Segundo, se aborda el LLM como un modelo de IA que tiene la capacidad de reconocer y entender el lenguaje natural, esto es planteado por González San José (2025) en su estudio *Análisis de flexibilidad de los modelos LLM cuantizados.* Para el autor, los LLMs utilizan el contexto para la producción de significados, generando respuestas con relevancia y coherencia a partir de mecanismos de predicción. Aquí, la definición da una mirada a la revisión del contexto para la comprensión del lenguaje, pero no se enuncia desde el lugar del sesgo de entrenamiento, por lo que no es compatible con el objetivo conceptual de esta investigación.

Finalmente, en el trabajo *La privacidad de los LLM en la enseñanza obligatoria,* escrito por Héctor Poveda (2024), se hace referencia a los LLMs como grandes modelos de lenguaje que se entrenan en grandes corpus de datos a partir de la información encontrada en internet, literatura y demás medios de información. Aquí se pone en cuestión la generación de respuestas ya que se establece que la generación de contenido puede ser errada o basada en sesgos que se encuentran presentes en los datos de entrenamiento. Esta definición que se sitúa de manera crítica en las formas de entrenamiento es la que permite abordar el problema planteado dentro de este escrito y el abordaje del caso de sesgo frente a culturas locales.

Tokenización

Para aproximarse a este concepto, es importante comprender que tokenizar implica una transición entre el flujo continuo del lenguaje natural y las estructuras inteligibles por las arquitecturas de aprendizaje profundo, allí se da un proceso de segmentación y codificación conocido como tokenización. Empero, este no es un paso netamente instrumental o neutral, aquí se delimitan las fronteras de los distintos significados de las palabras para los modelos. Por ello, el alcance del token se debe articular en tres dimensiones interconectadas; la lingüística, la computacional y la metodológica. En este sentido, teóricamente se hace

necesario dar una mirada a las distintas formas de abordar la tokenización, seleccionando la que mejor se adapte al proceso investigativo.

En primer lugar, se encuentra la percepción de Ali et al. (2024) en el trabajo *Tokenizer Choice For LLM Training: Negligible or Crucial?,* aquí la tokenización es definida como un método básico que "...consiste en dividir las secuencias en función de los espacios en blanco y considerar cada palabra como un token" p. 3908. Dicha precisión sobre este proceso no concuerda con el objetivo buscado dentro de este trabajo, puesto que sólo propone una forma de realizar la tokenización. Es necesario comprender que existe diversidad de formas de tokenizar y que esto es lo que enriquece el proceso en pro de la visibilidad de escenarios no dominantes.

Para un segundo abordaje se encuentra la investigación de Ardila (2023), que establece la tokenización como modelos lingüísticos "... que funcionan dividiendo el texto en elementos más pequeños, llamados tokens, para facilitar su análisis y predicción. Un token puede ser un solo carácter o también una palabra, esto depende de la tarea que se busque" p. 22. Esto que el autor propone como tokenización, tiene un contexto más amplio de cómo funciona el proceso, pero no se ajusta a la delimitación del problema ya que no aborda la diversidad lingüística de una manera que se puedan visibilizar otras culturas.

Finalmente, el recorrido conceptual se detiene en la definición concebida por Sánchez (2025) en su trabajo *Generación automática de recursos de aprendizaje utilizando Large Language Models (LLM).* Aquí, el autor plantea la tokenización como "...la división de un texto en tokens, y este proceso puede ayudar al modelo a manejar diferentes idiomas, vocabularios y formatos, y a reducir los costos computacionales y de memoria" p. 44. Dicha percepción se retoma para el desarrollo del presente trabajo puesto que, su diálogo con el objetivo de visibilizar la diversidad lingüística y cultural se evidencia al entender el concepto como un proceso plural y sostenible.

Soberanía tecnológica

Hacer referencia a la soberanía tecnológica implica pensar en la autonomía de los países frente a los avances tecnológicos. Para esta definición se acude a tres autoras (es) que tienen un punto de encuentro: la democratización de las estructuras digitales. Primero, se encuentra

el trabajo de Ceballos et al., (2020) titulado *Soberanía tecnológica digital en Latinoamérica.* Los autores establecen que la soberanía es "una estrategia adoptada en función de los entornos en los que se desempeña." p. 158. Los autores, sostienen que este concepto ha sido transitado desde la esfera estatal hasta el terreno de las TIC's. De esta manera sostienen que "en tiempos de big data, sin una democratización y transparencia en los mecanismos de recaudación, vigilancia y control de la información y el conocimiento se provoca una erosión notable de la soberanía estatal y popular" (Ávila Pinto, 2018 en Ceballos et al., 2020, p. 152). Esto implica una mayor demanda en materia de tecnología que sea accesible para la población y se acerca un poco al objetivo del análisis planteado en este informe. Sin embargo, es necesario ampliar la definición buscando la mayor compatibilidad con el caso propuesto.

En un segundo momento, en el artículo *Inteligencia Artificial y Desigualdad Tecnológica Global: Impactos en la Soberanía Digital de Países en Desarrollo* -escrito por Elefante Books & Educational Technologies (2025)-, se entiende la soberanía tecnológica como aquello que sucede "al importar modelos fundacionales como los LLM [...]—que han sido optimizados para maximizar funciones de utilidad definidas en Silicon Valley o Shenzhen—, las naciones receptoras importan también los sesgos, valores y visiones del mundo implícitos en esos modelos" p. 43. Dicha forma de explicar lo que representa la soberanía tecnológica habla sobre las formas en las que el sesgo se sostiene a través de la transferencia de información de un país a otro, sin contextualizar el modelo. Esta es una definición que sirve para reflexionar de manera crítica sobre los cambios que se deben implementar en estos escenarios de acceso a la información, para entrenar modelos más incluyentes y accesibles.

En tercer y último lugar, las autoras Morales et. al (2025) plantean una percepción de la soberanía tecnológica basada en casos de Latinoamérica, donde la soberanía tecnológica "... abarca un espectro más amplio, centrándose en la autonomía política de los pueblos sobre sus infraestructuras digitales, la gobernanza de datos y las condiciones de acceso y desatando la dependencia de corporaciones transnacionales y el colonialismo de datos " p.12. Su artículo *Tecnopolíticas feministas: resistencias digitales, justicia de datos y soberanía tecnológica en América Latina* permite reflexionar sobre las infraestructuras digitales autónomas y comunitarias.

Este escenario da una mirada con un enfoque que se acopla más al análisis propuesto ya que percibe las infraestructuras mencionadas "...como un ámbito estratégico para disputar

el acceso, el control y la soberanía tecnológica en territorios históricamente marginados" (Morales et al., 2025, p. 9). Adicionalmente, este trabajo plantea una mirada frente a la carencia económica y de cobertura que tienen algunos territorios, dejando la concentración del mercado en manos de pocos actores que pertenecen al sector privado. Por ello, es importante la creación de redes comunitarias que ayuden a democratizar los modelos (Morales et al., 2025).

Triangulaciones y análisis. Google DeepMind: Al Research Foundations.

En este apartado se busca presentar la descripción y análisis de los contenidos del curso *Google DeepMind: AI Research Foundations,* presentando de manera secuencial las condiciones teóricas y técnicas de cada sección del curso. Aquí, no solo se presenta una síntesis descriptiva de los módulos, sino que se da una contrastación y reflexión teórica con el objetivo de revisar la aplicabilidad técnica de cada uno de los bloques de conocimiento proporcionados por el curso. De esta manera, será posible identificar cómo las arquitecturas y metodologías de entrenamiento logran resolver —o sostener— la invisibilización de la diversidad cultural e identitaria y el sesgo del Norte Global. Dicho recorrido analítico se realizará en una división de 6 módulos según la ruta establecida por el proceso de aprendizaje de Google.

Módulo 1: Build Your Own Small Language Model.

Para comenzar, este momento introductorio del curso permitió examinar cuatro aspectos esenciales para la modelación lingüística. En primer lugar, se pusieron a disposición los fundamentos de predicción, donde se establece que un modelo de lenguaje es un sistema que tiene la posibilidad de calcular la probabilidad de cuál es la palabra que sigue después de la que antecede. Estas palabras, concebidas como tokens, son las que se deben pluralizar al momento de entrenar los modelos de lenguaje. Si el modelo predice la siguiente palabra basado sólo en probabilidades estadísticas, habrá un vacío en las lenguas y modismos, que en este caso son colombianos. Esto, debido a que la visibilidad de los modismos culturales de Colombia aparece menos en los datos globales.

Tal y como lo plantea Gómez (2025) una poca variedad de palabras en la salida de un modelo suele deberse al "pequeño tamaño del corpus" de entrenamiento. El autor argumenta que solo con más datos es posible observar más palabras y obtener estimaciones de probabilidad más precisas. Es decir, un modelo del lenguaje colombiano tiene una visión más limitada y menos precisa a comparación del español peninsular o el inglés.

En segundo lugar, se aborda la evolución técnica a partir de una contextualización lineal que permite contrastar los modelos antiguos (n-grams), que fallaban por falta de datos (data sparsity) con los Transformers, que pueden entender grandes volúmenes de información. Aquí, es importante dar una mirada crítica desde el caso propuesto a evaluar puesto que, si bien el modelo Transformer representa un avance, aún se encuentran en deuda con el significado real de conceptos ancestrales. Esto se da si el Transformer no tuvo un entrenamiento con la profundidad necesaria, dicha capacidad de tener contextos más amplios de información se queda corta cuando el data set no cuenta con una información cultural completa.

En un tercer momento del módulo, se aborda el Pipeline de Machine Learning donde se estableció el paso a paso para crear una IA. Este espacio, brindó las herramientas propicias para el proceso de limpieza de textos (tokenización) y el entrenamiento de los modelos para reducir las posibilidades de error (loss). Teniendo en cuenta la diversidad cultural que se enuncia en el caso propuesto, son estos procesos de limpieza de textos los que convierten las identidades, prácticas y lenguajes culturales en números. Si el proceso de tokenización es deficiente, la cultura queda invisibilizada. De esta manera será importante que, para reducir el sesgo en el entrenamiento de IA, se realicen protocolos de tokenización que puedan abarcar la pluralidad cultural de acuerdo con la geolocalización.

Finalmente, esta sección del curso brinda la posibilidad de tener una mirad apor fuera de la visión técnica a través de la ética situada (Ubuntu). Aquí se plantean valores comunitarios a través de conceptos como el Ubuntu (enfoque en la comunidad) cuestionando de manera crítica si la IA sirve a las personas o solo a las empresas. Este cierre del módulo permitió pensar en las formas en las que el entrenamiento de Transformers afecta la diversidad cultural y las comunidades. Es un buen enfoque para pensar en las formas en las que el capital cultural se sitúa sólo en ciertas esferas poblacionales, dejando de lado la multiplicidad de identidades y comunidades que existen en el medio.

Módulo 2: Represent Your Language Data

Este segundo módulo, representa uno de los momentos más importantes dentro del procesamiento de lenguaje: la conversión de un texto en representaciones numéricas que pueda leer una máquina. Aquí, se da la codificación del lenguaje y la estructuración o exclusión de diferentes formas de ver. Para iniciar, se aborda el componente técnico de preprocesamiento de texto real; en este espacio se aprende a limpiar los datos considerados como sucios (quitar etiquetas HTML, manejar caracteres especiales) para que el modelo reciba únicamente la señal importante.

Posteriormente, se enfoca la discusión en la tokenización de subpalabras (BPE); donde es posible observar por qué realizar segmentaciones por palabras completas o por letras no funciona bien. Este apartado del módulo 2 posibilitó la implementación del algoritmo Byte-Pair Encoding (BPE), el cual divide las palabras en fragmentos comunes, aspecto que permite que el modelo emplee palabras que no estaban en su diccionario original. Con respecto al caso propuesto, esta óptica del algoritmo BPE permite revisar el entrenamiento de las redes neuronales.

Si el BPE es entrenado con poco texto del lenguaje colombiano las palabras de la identidad del país se fragmentarán hasta perder la conexión semántica, haciendo que la cultura local tenga un costo incrementado computacionalmente y que sea menos precisa. Esto en palabras de Lundin et al. (2026), "...cuantificamos el impacto económico de la ineficiencia de la tokenización, demostrando cómo el "impuesto al token" crea barreras para el desarrollo del PLN multilingüe" p. 104. Es decir, donde el sesgo de tokenización termina siendo pagado de manera desproporcionada por los hablantes de lenguas morfológicamente complejas y con pocos recursos. Aspecto que se convierte en un modelo insostenible para países del tercer mundo, pero que también aísla ciertos sectores poblacionales del medio. Para fortalecer esta postura, es importante remitirse a los ejemplos del curso. Según Google (2026) "la mayoría de los modelos preentrenados se basan principalmente en textos en inglés. Sus tokenizadores están altamente optimizados para el inglés, pero pueden resultar ineficientes para muchos otros idiomas, incluidos los idiomas africanos." S.p.

En la tercera parte del módulo 2, se dirige el diálogo hacia los vectores de significado (embeddings). Allí, se establece que cada token se convierte en un vector numérico que hace parte de la creación de un mapa de significado donde la distancia entre palabras indica qué tan relacionadas están. Dicho mapa de significados implica tener en cuenta que, si los vectores se agrupan según valores ajenos, las palabras de morfologías como la colombiana quedarían excluidas o muy lejos de los conceptos importantes. Esta es una forma matemática de representar la marginación y el sesgo mencionado dentro del análisis.

Adicionalmente, se recurre a las técnicas t-SNE y a la similitud del coseno[3] para la visualización de gráficos 2D donde puede evidenciar con métricas las formas en las que el modelo agrupa los conceptos. Estas herramientas de visualización permiten al arquitecto de datos, observar si el modelo malinterpreta o agrupa de manera errada los conceptos de la cultura local. Esta función actúa en los casos de estudio como una auditoría visual ante el sesgo.

En último término, para esta fase del curso se analiza la importancia de los datos a través de un estudio de caso. Esto, como una forma de relacionar los data sets y la ética. Dicho caso plantea cómo la falta de representación en los data sets genera exclusión digital y de manera crítica cuestiona por qué se necesitan herramientas de transparencia como las Data Cards. En este escenario, la soberanía tecnológica implica tener hojas de datos transparentes en modelos usados en Colombia para permitir la declaración de qué comunidades fueron incluidas y cuáles fueron ignoradas en el entrenamiento de las redes neuronales. Aquí, se hace necesario comprender que las tecnologías digitales relacionadas con la IA y el big data están concentradas en un oligo-polio donde se comienza a establecer un único capital cultural proveniente del norte global y se designan las reglas de la economía global digital (Ortíz de Zárate, 2024). De esta manera, el autor plantea que

---

[3] Esta técnica "mide el ángulo entre dos vectores y proporciona una puntuación entre -1 y 1. Una puntuación de 1 significa que los vectores apuntan exactamente en la misma dirección (son perfectamente similares). Una puntuación de 0 significa que los vectores forman un ángulo de 90 grados (no están relacionados). Una puntuación de -1 significa que apuntan en direcciones opuestas (son opuestos)." (Google, 2026, s.p). Es posible calcular la similitud del coseno de dos vectores de n dimensiones en el idioma de mejor entendimiento. Es la herramienta que usa la IA para medir qué tan parecido es el significado de dos textos.

> ...datos sensibles, procesos nacionales sociopolíticos y económicos fundamentales e infraestructura crítica quedan a cargo de empresas extranjeras que responden indudablemente a sus intereses y a los de su Estado de origen. La dominación en esta materia conduce a un colonialismo tecnológico (Ávila Pinto, 2018 en Ortíz de Zárate, 2024, p. 159).

Por lo que, como se propone en el caso planteado, es importante democratizar, visibilizar y tener transparencia en las formas de recolección de información para el entrenamiento de las redes neuronales en regiones como América Latina.

Módulo 3: Design And Train Neural Networks

Este tercer momento del programa de formación de Google traslada el enfoque de la representación de datos hacia los mecanismos internos que posibilitan que las redes neuronales puedan aprender y parametrizar la información. De acuerdo con el curso

> El objetivo es lograr que el sistema sea lo suficientemente inteligente como para manejar situaciones nuevas que nunca antes ha visto. Una buena red puede tomar lo que aprendió de sus ejemplos de práctica y aplicar ese conocimiento a una nueva oración que nunca antes ha visto (Google, 2026, s.p)

Aquí se evidencia que el estudio del diseño y el entrenamiento de las estructuras neuronales supone una inmersión en el motor de aprendizaje profundo. Este es un escenario donde las decisiones que definen la capacidad del modelo para poder interpretar la realidad son matemáticas. Es importante tener en cuenta que este no es un proceso exclusivamente técnico, sino que es un plano donde se pueden estabilizar los sesgos culturales de origen.

En un primer momento, se plantea que el éxito no depende solo de la capacidad del modelo para memorizar los datos de entrenamiento, sino que este pueda actuar ante datos que no había reconocido antes. Se hace necesario plantear que, si el modelo solo generaliza los datos para el español estándar o contextos urbanos, es porque el universo de datos al que ha tenido acceso es limitado. Esto logra evidenciar que la constante presencia de errores en

la generalización de culturas como la colombiana no se instaura en el proceso, sino en la ausencia de exposición del modelo en entrenamiento a la diversidad local. De acuerdo con el curso

> Esto implica reflexionar detenidamente sobre cómo podrían verse afectadas las comunidades, las prácticas culturales y los flujos de trabajo organizativos al introducir un modelo de aprendizaje local (LMM). Por ejemplo, la implementación de una herramienta de planificación urbana basada en un LLM podría mejorar la participación ciudadana, pero también podría alterar la dinámica de poder entre los residentes y los funcionarios municipales, modificar la forma en que se toman las decisiones y, potencialmente, relegar conocimientos locales importantes o redes comunitarias informales (Google, 2026, s.p).

Posteriormente se da una mirada comparativa al sobreajuste (Overfitting)[4] y el subajuste (Underfitting). En este caso, se aborda el riesgo al que se accede cuando un modelo es muy simple y no capta gran cantidad de información o muy complejo y alcanza a memorizar el ruido o sesgos de los datos, por lo que termina fallando en la vida real. Actualmente, los modelos tienen una tendencia al sobreajuste cultural, donde se han memorizado y replicado de manera masiva los patrones del norte global. Este escenario, evidencia que los modelos han perdido la capacidad de entender patrones lingüísticos que se encuentran por fuera del estándar, tal y como sucede con la cultura colombiana.

En un tercer momento, se abordan las afecciones a la capacidad del modelo para aprender patrones. Esto depende del número de capas y neuronas que se implementen al momento del entrenamiento. Adicionalmente, se logra entender cómo este factor, convierte al modelo en un escenario propenso al error si no es bien gestionado. Es decir, mientras más

---

[4] Existen distintas formas de mitigar el overfitting, esta es fundamental para desarrollar modelos de aprendizaje automático robustos que se generalicen bien a nuevos datos. Esto, a partir de estrategias como el control de capacidad, la regularización (decaimiento de pesos), el abandono (dropout), los puntos de control y la detención temprana, se pueden construir modelos precisos y fiables (Google, 2026, s.p)

complejidad tenga el modelo, más ruido aprenderá; por ello, es importante contar con una guía ética para que este pueda aprender y expandir los significados reales en lugar de reproducir de firma masiva los estereotipos culturales o ruidos sociales.

Por otra parte, el cuarto momento del módulo se enfoca en las técnicas de control a partir de la implementación de herramientas como Weight Decay, Dropout y Early Stopping para direccionar al modelo en vía de ser más robusto y menos apático o dependiente de ciertos datos. Aquí, es importante reflexionar frente a las situaciones que el modelo considera un error; si el modelo es castigado (vía backpropagation) por no tener un pronóstico de la palabra en español estándar, quiere decir que el sistema está siendo entrenado constantemente para suprimir o invisibilizar los modismos locales en favoreciendo la precisión estadística.

Además, se dio una mirada al motor del aprendizaje del modelo, evidenciando cómo la retro-propagación (backpropagation)[5] y el descenso de gradiente estocástico (SGD) son los algoritmos que dan equilibrio a los pesos del modelo para reducir el error. De acuerdo con el curso

> los gradientes desempeñan un papel fundamental en el entrenamiento de redes neuronales. En lugar de ajustar manualmente los pesos, permiten que el modelo aprenda automáticamente. La idea principal es que el gradiente indica cómo modificar los pesos para reducir la pérdida (Google, 2026, s.p)

Esta técnica, conocida como anticipación tiene relación con la coyuntura local planteada para el ejercicio. Es entonces donde el arquitecto de datos no sólo debe prever la caída del servidor, sino pronosticar la manera en la que el modelo puede perpetrar desigualdades existentes o -en su defecto- invisibilizar una lengua al tener interacción con otras culturas como la colombiana.

---

[5] "El mecanismo central de la retro propagación consta de un ciclo de dos etapas: primero, la pasada hacia adelante, donde los datos de entrada se procesan a través de las capas de la red para generar una predicción; segundo, la pasada hacia atrás, donde el algoritmo trabaja a la inversa desde la función de pérdida para calcular eficientemente el gradiente, o contribución de error, para cada peso y sesgo en la red. El descenso de gradiente estocástico utiliza estos gradientes calculados para actualizar los parámetros del modelo de manera que se minimice la pérdida." (Google, 2026, s.p)

Para el cierre de este módulo se introdujo la reflexión sobre los procesos de investigación en el campo. Esto, porque se considera que el ejercicio investigativo no sólo implica prever fallos técnicos, sino que conlleva a anticipar cómo la IA va a interactuar con contextos sociales y culturales complejos. En relación con el proceso formativo de Google (2026) es importante "identificar riesgos potenciales, como sesgos, exclusión o violaciones de la privacidad. También implica considerar cómo la innovación podría afectar valores sociales fundamentales como la equidad, la dignidad, la autonomía y el bienestar comunitario" s.p., la anticipación ayuda a prevenir daños potenciales especialmente en grupos marginados que pueden verse afectados de manera desproporcionada (Google, 2026). Aspecto que indica que la creación de las inteligencias artificiales responde a un acto de diseño colectivo que es definido por Constanza-Chock (2018) como un elemento "clave para nuestra liberación colectiva, pero la mayoría de los procesos de diseño actuales reproducen desigualdades" p. 530. Puesto que la confiabilidad de un modelo en territorios como el colombiano no se puede medir únicamente por su baja pérdida (loss), sino por la alineación frente a las realidades y necesidades de la colectividad.

Módulo 4: Discover The Transformer Architecture

Este módulo muestra una ruptura metodológica con respecto a los bloques anterior o secuencial que se utiliza tradicionalmente se ve desplazado por las dinámicas de paralelización masiva y asignación de probabilidad relevante. Dicho proceso, ha posibilitado observar la transferencia de conocimiento técnico avanzado del curso, donde se aborda la arquitectura transformer como el pilar fundamental sobre el cual se edifican los grandes modelos de lenguaje (LLM). Empero, la evaluación técnica de estas arquitecturas se convierte en un complemento para comprender las diferentes dimensiones –como la sociológica- que impactan la automatización en las interacciones humanas.

La primera parte del módulo hace referencia a las decisiones del modelo a partir del mecanismo de atención y multi-head attention, con esta técnica fue posible entender cómo el modelo decide a qué partes de una frase prestar atención. Aquí, la multiplicidad de cabezas de atención permite que el modelo vea diferentes aspectos al mismo tiempo como la gramática, los sentimientos o el contexto. Teniendo en cuenta que este curso está siendo

analizado bajo las posibilidades de acción frente al sesgo de la IA en regiones por fuera del norte global como Colombia, es importante mencionar que la atención técnica es una forma de jerarquización cultural. En este sentido, el modelo puede o no aprender a prestar atención a los marcadores de lenguaje colombianos; si este no lo aprende, estos se dejan de lado como información secundaria ante estructuras globales.

De acuerdo con lo anterior y la continuidad del curso, la segunda sección del módulo 4 hace referencia al attention mask y positional embeddings; presentando las formas en las que se le puede enseñar al modelo el orden de las palabras. Esto, porque los Transformers procesan todo en paralelo y, sin esto, no sabrían qué palabra va primero. Es conveniente mencionar que, en el caso de las lenguas locales, el orden de las palabras puede variar puesto que tienen un orden propio. Por ello, es importante tener una mirada crítica ante el uso de los embeddings posicionales diseñados para las lenguas dominantes, ya que esto puede forzar estructuras de pensamiento ajenas sobre las expresiones nativas o regionales.

En un tercer momento, el curso alude a las estructuras internas que incluyen la normalización de capas –para la estabilidad- y el MLP que añade complejidad y no linealidad en el bloque transformer. Esta parte del módulo conlleva a una discusión importante sobre la búsqueda de estabilidad y normalización; la técnica en el bloque transformer contribuye a una homogeneización del lenguaje. Aspecto por el cual, se da la eliminación de las particularidades de identidad dejando de lado lenguajes más complejos e invisibilizados que están por fuera del norte global.

De allí, que desde el mismo módulo se hable después sobre las prácticas sociales y su relación con la automatización. Esto, permite reflexionar frente a cómo las interacciones cotidianas del lenguaje -como saludar- no son solo datos, sino actos que crean confianza y tejido social. De acuerdo con el curso

> los saludos y las conversaciones se aprenden y se practican colectivamente, dando forma a los ritmos de la vida comunitaria. Los niños crecen observando que, incluso en los negocios, la medicina y el trabajo, las relaciones importan. Así, las interacciones cotidianas se convierten en prácticas pequeñas pero vitales que dan sentido a las cosas y que mantienen los lazos sociales.

Por ello, se advierte que –de acuerdo con las formas de interacción locales- cuando se automatizan las interacciones sociales propias de Colombia donde los vínculos humanos son un factor clave; pueden generarse pérdidas de capital cultural y social si la IA comienza a reemplazar estas prácticas en el lenguaje por una eficiencia que sea puramente numérica.

Finalmente, en el módulo 4 se establecen las formas de identificación de quienes se ven afectados por los avances tecnológicos como las comunidades indígenas, los trabajadores y el sector académico. Esto se hace bajo la técnica mapeo de stakeholders y plan de engagement donde no sólo se rastrea quienes se ven afectados, sino que se da paso a la búsqueda de inclusión en el diseño de los modelos. Metodológicamente, y teniendo en cuenta el caso propuesto, estos diseños de LLMs principales deben incluir de manera obligatoria un mapeo de actores locales, en este sentido es pertinente aproximarse a lo planteado por Google (2026), donde se establece que es importante "reconocer que las tecnologías afectan a los grupos de maneras muy distintas, según sus valores, contextos y posición en la sociedad. Sin una participación significativa, los desarrolladores corren el riesgo de incurrir en aumentar las desigualdades" s.p.

Aquí, es importante recurrir a formas de innovación responsable, exigiendo que el modelo sea construido con los grupos que serán representados y no solo para el consumo de estos. De esta manera, se extienden las garantías de visibilización de las comunidades sin excluir la cultura, la identidad y el lenguaje en los procesos de entrenamiento de redes neuronales.

Módulo 5: Fine-Tune Your Model

El quinto ápice del curso permite transitar de manera crítica de los modelos funcionales de IA hacia su especialización y despliegue situado. Este módulo plantea que, si bien un modelo base tiene capacidades lingüísticas generales, hay una carencia en la contextualización necesaria para operar con pertinencia cultural y ética en escenarios como el propuesto para este análisis, con especificidades identitarias y culturales que no son parte del estándar. Es un bloque en el que se trasciende la discusión desde la infraestructura masiva hacia metodologías más amables que promueven la adaptación técnica y la alineación. Esto

es importante porque se propone el ajuste fino, no solo como procedimiento de optimización, sino como un espacio de gobernanza.

En ese sentido, se estructura el módulo en 6 puntos clave. Primero, con una aproximación al Ajuste Fino Supervisado (SFT) se identifica que los modelos base (como Gemma) tienen capacidades generales, pero el Ajuste Fino Supervisado logra una adaptación de los modelos a tareas específicas a partir del reentrenamiento en data sets pequeños de alta calidad. Para el análisis, esta técnica es una vía para la soberanía tecnológica porque no es necesario crear un modelo desde cero; basta con tomar uno global y enseñarle las particularidades lingüísticas y culturales de regiones como Colombia a partir de data sets locales de alta calidad.

Segundo, a partir de la eficiencia con LoRA (PEFT) se evidencia que re-entrenar todo un modelo es computacionalmente fuera de acceso para la mayoría, pero con el uso de LoRA –que congela el modelo original y solo entrena unas pocas capas nuevas– (matrices de bajo rango) se logra hacer una adaptación accesible y eficiente. Esta forma de acción para el entrenamiento permite acceder a la viabilidad económica, donde el arquitecto de datos puede lograr una representación cultural sin presupuestos de grandes multinacionales. De esta manera, es posible una democratización en la capacidad de intervenir los modelos a partir de técnicas de bajo rango.

Tercero, se da una exploración frente a cómo el feedback humano tiene una incidencia importante sobre el modelo, no sólo para que sea preciso, sino para que pueda estar alineado con valores humanos, seguridad y utilidad; más allá de las predicciones estadísticas. Esta técnica se conoce como Aprendizaje por Refuerzo (RLHF) y frente al análisis propuesto para este informe, se debe alinear el modelo con un feedback de cada región -en este caso Colombia-, puesto que, si el refuerzo humano sólo proviene del norte global, el modelo continuará con una exclusión de los valores regionales. En esta medida, se hace necesaria la implementación de RLHF que sea situado de acuerdo con cada territorio.

Cuarto, el módulo 5 dispone una reflexión desde el significado cultural y las narrativas, cuestionado sobre cómo prácticas como la enseñanza o el saludo se encuentran cargadas de valores establecidos desde la ética. Aspecto que supone el peligro de que la IA ignore estos significados y socave la confianza o refuerce la desigualdad estructural. Por ello, como se ha mencionado en apartados anteriores, existe el riesgo ético de que las narrativas

locales sean invisibles a los modelos. Estos, deben tener la capacidad de comprender que el lenguaje en cada una de las regiones latinoamericanas es un vehículo de identidad, pertenencia y autoridad, por lo que no debe percibirse sólo como un intercambio de información.

Ante el panorama anterior, desde el proceso de aprendizaje se plantean dos escenarios: el de la gobernanza y la prospectiva, pero también el de la responsabilidad y los futuros colectivos. Con respecto al primer escenario, se plantea que la gobernanza no puede ser solo para las élites o las instituciones. Esta debe integrar valores culturales, reglas claras y la exigencia de pruebas que indiquen que los sistemas son beneficiosos antes de ser desplegados. El segundo panorama que corresponde al último apartado de este módulo permite concluir que, si los LLM van a tener incidencia en los cambios de la sociedad, esta misma es la que debe tomar la decisión de cómo se hace. Así pues, la IA debe ser modelada con participación colectiva y pública, con valores compartidos y no únicamente por diseño técnico.

En este sentido, es necesario hablar sobre la gobernanza democrática de la IA como una aplicación práctica a la soberanía tecnológica digital, partiendo de una exigencia a que las comunidades tengan voz en las reglas de despliegue de los modelos. De igual forma, la consideración del diseño de LLM como un acto de deliberación pública implica que el arquitecto de datos no solo se dedique a la optimización de los códigos, sino que asegure que la IA refleje un futuro colectivo en regiones como Colombia, protegiendo la diversidad cultural.

Módulo 6: Accelerate Your Model

Para dar cierre al curso, el último módulo se enfoca en el giro crítico del diseño algorítmico abstracto hacia las restricciones físicas y materiales que condicionan el despliegue de la IA. En este apartado, se plantea que el modelo de aprendizaje profundo depende de una infraestructura de hardware, energía y recursos naturales para su existencia, independientemente de lo complejo que este pueda ser. Según el curso

> los sistemas de aprendizaje automático (LLM) consumen muchos recursos tanto durante la formación como en su uso diario. Los centros de datos que

> los gestionan consumen mucha energía y requieren una amplia infraestructura de refrigeración, que a su vez consume electricidad y agua. Este importante consumo de recursos genera una considerable huella de carbono y otras emisiones contaminantes (Google, 2026, s.p.).

Por ello, el módulo final establece una reflexión donde estudiar los límites de cómputo y las técnicas de aceleración dan la posibilidad de comprender que la soberanía tecnológica no sólo es una postura epistémica, sino que también representa un desafío económico y material que se condiciona por la desigualdad global de acceso al hardware.

Así pues, se da inicio al primer apartado del módulo final[6] donde -a partir del Trade.off entre rendimiento y eficiencia- se analizan los modelos más grandes como Gemma-4B y cómo estos ofrecen una mayor calidad. Sin embargo, que se aclara el costo que debe ser asumido desde estos modelos, es más alto en tiempo y recursos que los modelos pequeños. Según Lundin et al., (2026) "Un costo razonable para entrenar un modelo pequeño-mediano o un modelo de frontera grande oscila fácilmente entre $1 millón (1 mes) y $100 millones (aproximadamente 3 meses) con tokens principalmente en inglés" p. 103. Para el caso de la soberanía tecnológica en Colombia, es más accesible un modelo más pequeño y eficiente que uno masivo y lento. La eficiencia permite tener autonomía técnica.

Se da continuidad al módulo con las mediciones del costo computacional mediante FLOPs y el cálculo de la memoria necesaria a la GPU para parámetros, datos, gradientes y estados del optimizador. Aquí, se abordaron los recursos críticos que recogen el cómputo y la memoria, donde el arquitecto de datos debe ser consciente de las limitaciones de infraestructura en el contexto local. Tal y como lo plantean Elefante Books & Educational Technologies (2025) "la IA moderna es computacionalmente costosa. Entrenar los modelos de frontera requiere super computadores y centros de datos hiperescalares que consumen la energía equivalente a pequeñas ciudades" p. 39. En este sentido, no se trata sólo de correr un código, sino de optimizar el hardware disponible en la región.

De la misma forma, se trabaja el muro de la memoria (memory wall) para explorar los errores técnicos que ocurren donde se intenta entrenar modelos grandes con hardware

[6] Es necesario clarificar que el módulo 6 no existe, el curso no lo hace visible y salta del módulo 5 al 7.

estándar y la manera de diagnosticarlos. En este escenario, el muro de la memoria se presenta como una barrera de entrada; superarlo con técnicas como bfloat16 es lo que permitirá que países con menos recursos de cómputo puedan participar en la creación de IA. Es decir, las técnicas exploradas en este apartado del curso favorecen la democratización tecnológica y la reducción del sesgo.

Adicionalmente, desde las técnicas de aceleración y optimización fue posible aproximarse a la implementación de soluciones como el uso de herramientas bfloat16. Con estas, es posible reducir el uso de memoria por parámetro y la acumular gradientes para simular lotes grandes sin saturar la memoria. Si se da uso a estos métodos de optimización se generan técnicas y estrategias para que la IA sea más accesible y menos dependiente de instrumentos de alto costo como las super computadoras utilizadas por las grandes empresas del norte global.

El cierre de este módulo y del curso en general se da de manera reflexiva ante el vínculo material de la IA. Aquí, es importante tener conciencia de que esta supone un consumo masivo de energía, agua, labor y hardware. Aspecto que implica plantear preguntas urgentes con una mirada crítica sobre quién asume los costos ambientales. Por ello, dentro de este informe se recurre a la soberanía tecnológica y se sugiere que esta debe ser ecológica, por fuera de los modelos globales que consumen recursos locales como la energía o el agua para los servidores sin dejar un beneficio claro para el territorio colombiano y su biodiversidad. Adicionalmente, desde enfoques como la justicia y la sostenibilidad, este proceso de formación cierra con un mapeo hacia la huella de carbono y la idea de una transición energética justa. Aquí, se propone que los beneficios de la IA no estén enfocados en la extracción de recursos que dañen el bienestar local.

Conclusiones

El contenido conceptual y analítico que se desarrolló dentro de este informe demuestra que el LLM es una herramienta con gran potencial y poderosa. Sin embargo, puede resultar ser un arma de doble filo dependiendo del manejo que se le dé. Si bien los LLMs pueden ayudar a la visibilización de identidades y culturas deberán tener buenas prácticas al

momento de entrenar los modelos; de no ser así, estos pueden terminar desapareciendo a las comunidades que no hacen parte del capital cultural perteneciente al norte global.

En este sentido, es posible afirmar que los modelos comerciales de IA sostienen una profunda desigualdad de representación que favorece de manera hegemónica a las perspectivas del norte global, convirtiendo su capital lingüístico y de datos en el único legitimado por la web. De esta manera, los mayores índices de sesgo y exclusión digital están concentrados en las periferias del mundo afectando de forma desproporcionada las culturas que no hacen parte del norte global, como las comunidades latinoamericanas y colombianas. Aquí, el borramiento digital se convierte en un ejercicio sistemático puesto que las expresiones locales, jergas y lenguas nativas o minoritarias son tratadas estadísticamente como anomalías por los modelos de frontera entrenados con datos que no han sido debidamente curados.

En suma, como respuesta a la pregunta de análisis propuesta en este trabajo, se considera que sí existe la posibilidad de cerrar la brecha entre los LLMs y regiones culturalmente diversas. Esto implica un trabajo en conjunto entre gobiernos, capital privado con interés social y un mapeo de saberes ancestrales, culturales y sociales para alimentar el flujo de información y el entrenamiento de la IA.

Además, desde la parte técnica, es importante tener presente que se debe operar con el sesgo algorítmico desde las capas arquitectónicas más profundas de las redes neuronales y los Transformers, por lo que no es un proceso superficial. De la misma manera, se debe considerar que los procesos de tokenización comunes se encuentran altamente optimizados para el inglés, fragmentando excesivamente los textos de lenguajes no dominantes. Este aspecto impone ineficiencias computacionales e incrementan los costos (impuestos al token), generando desigualdades a nivel local.

Finalmente, se logra identificar —según la bibliografía enunciada en este documento— que existe de manera latente una deuda ética y social durante el desarrollo de los LLM modernos. Esto, debido a las formas de entrenamiento que se han implementado desde su creación, donde la información y la data que se proporcionó a los modelos fue creada a partir de lenguajes estándar, dejando de lado los saberes y conocimientos de otras culturas. Empero, estos procesos se convierten en algo insuficiente si no se dan transformaciones metodológicas en el medio.

Por ello, es de gran importancia crear metodologías viables para mitigar el sesgo e incrementar el enriquecimiento del capital cultural sin incurrir en los altos costos de ejecución. Ante este panorama, es necesaria la especialización y la alineación local bajo técnicas como el Ajuste Fino Supervisado, implementando modelos base que sean abiertos y combinándoles con la optimización de parámetros de bajo rango. Por último, no se debe dejar de lado el Aprendizaje por Refuerzo con Retroalimentación Humana que se encuentre situado en los territorios, esto garantiza la participación de las comunidades en la curaduría ética de datos. En este sentido, la soberanía tecnológica, deberá conceptualizarse desde la autonomía del código pero también desde la optimización del hardware disponible, incluyendo la responsabilidad ecológica y social que defienda el porvenir de los futuros colectivos frente al colonialismo de datos.

Referencias

Abeliuk, A. y Gutiérrez, C. (2021). Historia y evolución de la inteligencia artificial. Revista Bits de Ciencia, (21), 14–21. https://revistasdex.uchile.cl/index.php/bits/article/view/2767/2700}

Ali, M., Fromm, M., Thellmann, K., Rutmann, R., Lübbering, M., Leveling, J., Klug, K., Ebert, J., Doll, N., Buschhoff, J. S., Jain, C., Weber, A. A., Jurkschat, L., Abdelwahab, H., John, C., Ortiz Suárez, P., Ostendorff, M., Weinbach, S., Sifa, R., … Flores-Herr, N. (2024). Tokenizer Choice For LLM Training: Negligible or Crucial? En K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3907–3924). Association for Computational Linguistics. https://aclanthology.org/2024.findings-naacl.247/

Arana, Carlos (2021) : Redes neuronales recurrentes: Análisis de los modelos especializados en datos secuenciales, Serie Documentos de Trabajo, No. 797, Universidad del Centro de Estudios Macroeconómicos de Argentina (UCEMA), Buenos Aires. https://www.econstor.eu/bitstream/10419/238422/1/797.pdf

Ardila Barbosa, D, Carrillo Aranda, D y Ladino Perdomo, V. (2023). DETEL Identificación de textos elaborados por LLM. Universidad de La Sabana. Disponible en: https://hdl.handle.net/10818/60271

Avaro, D. (2025). *Del Token al Token: una arqueología conceptual de la tokenización digital.* Isagoge Monografías. https://www.danteavaro.com/isagoge/monografias/monograf_2.pdf

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

Bourdieu, P. (1997). Capital cultural, escuela y espacio social (I. Jiménez, trad.). Siglo XXI.

Costanza-Chock, S. (2018) Justicia en el diseño: hacia un marco feminista interseccional para la teoría y la práctica del diseño, en Storni, C., Leahy, K., McMahon, M., Lloyd, P. y Bohemia, E. (eds.), El diseño como catalizador del cambio - Conferencia Internacional DRS 2018, 25-28 de junio, Limerick, Irlanda. https://doi.org/10.21606/drs.2018.679

Elefante Books & Educational Technologies (2025). Inteligencia Artificial y Desigualdad *Tecnológica Global: Impactos en la Soberanía Digital de Países en Desarrollo. Telematique*, 24(2), Pp. 37-49. https://ojs.urbe.edu/index.php/telematique/es/article/view/3900

Google. (2026). *Google DeepMind: AI Research Foundations* [Curso en línea]. Google Skills. https://www.skills.google/paths/3135

Gómez-Rodríguez, C. (2025). Grandes modelos de lenguaje: ¿de la predicción de palabras a la comprensión? En A. Alonso Betanzos, D. Peña, & P. Poncela (Eds.), *La Inteligencia Artificial hoy y sus aplicaciones con Big Data* (pp. 73–98). Funcas. https://www.funcas.es/wp-content/uploads/2025/02/04.-Gomez-Rodriguez-Carlos_1.pdf

González San José, P. (2025). *Análisis de flexibilidad de los modelos LLM cuantizados* [Trabajo de fin de grado, Universidad de Cantabria]. Repositorio Institucional de la Universidad de Cantabria (UCrea). https://hdl.handle.net/10902/36180

Henao Ciro, R. D., & Muñoz Gaviria, D. A. (2024). Pensar la tecnología más allá de los artefactos: aproximaciones desde la filosofía de la técnica. Praxis, Educación y Pedagogía, (13), e20215442. https://doi.org/10.25100/pep.v0i13.15442

IBM. (s.f.). *¿Qué es la similitud de coseno?* Think Topics. https://www.ibm.com/es-es/think/topics/cosine-similarity

Isasi Viñuela, P., & Galván León, I. M. (2004). Redes neuronales. Pearson Educación. https://www.academia.edu/15349367/Libro_Redes_neuronaes

Jessica M. Lundin, Ada Zhang, Nihal Karim, Hamza Louzan, Guohao Wei, David Ifeoluwa Adelani y Cody Carroll. 2026. El impuesto al token: sesgo sistemático en la tokenización multilingüe. En Actas del 7.º Taller sobre Procesamiento del Lenguaje Natural Africano (AfricaNLP 2026), páginas 103-112, Rabat, Marruecos. Asociación de Lingüística Computacional. https://aclanthology.org/2026.africanlp-main.10/

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2021). Causal inference technology for algorithmic fairness. In Proceedings of the 2021 ACM FAccT Conference (pp. 811–812). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Lundin, J. M., Zhang, A., Karim, N., Louzan, H., Wei, G., Adelani, D. I., & Carroll, C. (2026). The Token Tax: Systematic Bias in Multilingual Tokenization. En E. A. Chimoto, C. Lignos, S. Muhammad, I. Abdulmumin, C. Siro, & D. I. Adelani (Eds.), *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)* (pp. 103–112). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.africanlp-main.10

Mendoza-Pabón, J. A., & Agudelo-Ospina, A. (2024). Prácticas pedagógicas y su relación con el pensamiento computacional en docentes de básica primaria. Praxis, Educación y Pedagogía, (13), e20215442. https://doi.org/10.25100/praxis_edcacion.vi13.15442

Morales, S., Natansohn, L., Molina, S., (2025). *Tecnopolíticas feministas: resistencias digitales, justicia de datos y soberanía tecnológica en América Latina*. *Signo y Pensamiento*, *44*. https://revistas.javeriana.edu.co/files-articulos/SyP/44(2025)/6562797014/index.html

Ortiz de Zárate, J. M., Dias, J. M., Avenburg, A., & Gonzalez Quiroga, J. I. (2024). *Sesgos algorítmicos y representación social en los modelos de lenguaje generativo (LLM)*. Fundar. https://cdi.mecon.gob.ar/bases/docelec/az6612.pdf

Ortiz de Zárate, J. M. (2024). Desafíos metodológicos y políticos en la auditoría de sesgos en modelos de lenguaje generativo (LLM). *Propuestas para el Desarrollo*, (8), 53–74. https://propuestasparaeldesarrollo.com/index.php/ppd/article/view/108

Pérez Sampedro, S. (2024). Inteligencia artificial y ética: Una comparativa de modelos en el análisis de sentimientos y sesgos [Trabajo de fin de grado, Universidad Rey Juan Carlos]. BURJC Digital. https://hdl.handle.net/10115/38408

Poveda Sánchez, H. (2024). La privacidad de los LLM en la enseñanza obligatoria. [Trabajo fin de máster, Universidad Miguel Hernández]. España. https://dspace.umh.es/handle/11000/35542

Rodríguez Fernández, L., Fernández Mena, A. L., Torres Magaña, M. P., Rodríguez Magaña, M. A., & Rodríguez Fernández, M. A. (2024). Inteligencia artificial en la educación:

Modelo de lenguaje de gran tamaño (LLM) como recurso educativo. Revista Ipsumtec, 7(2), 157–164. https://doi.org/10.61117/ipsumtec.v7i2.321

Sánchez, H. (2025). *Generación automática de recursos de aprendizaje utilizando Large Language Models (LLM)* [Trabajo de fin de grado/máster, Universidad de Valladolid]. UVaDoc Repository. https://uvadoc.uva.es/handle/10324/78484

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. En I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 5998–6008). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vidal León, D. (2024). *El Rosco LLM* [Trabajo académico, Universidad Complutense de Madrid]. Docta Complutense. https://docta.ucm.es/entities/publication/10c75769-65a1-49e8-b0e6-46629d3ec97c