Simulación del Consumo Energético de Modelos de Inteligencia Artificial en el Marco de la Transición Energética - SIMIAE

Samuel David Cely Quesada

Carlos Andrés Moreno Jaimes

Trabajo de Grado para Optar el Título de Ingeniero electricista

Director
Johann Farith Petit Suárez
Doctor en ingeniería eléctrica electrónica y automática

Codirector
Oscar Arnulfo Quiroga Quiroga
Doctor en tecnología

Universidad Industrial de Santander
Facultad de Ingenierías fisicomecánicas
Escuela de Ingeniería eléctrica, electrónica y comunicaciones
Bucaramanga
2026

Dedicatoria

A mis padres, Mariela Quesada y Samuel Cely, por siempre creer en mí, por materializar este logro, por estar a mi lado en cada paso y ser mi ejemplo de vida, esto es gracias a ustedes.

A mis hermanos, Franky Cely y Diana Cely, por apoyarme incondicionalmente, por enseñarme a avanzar y no rendirme, por sus consejos y paciencia, por ayudarme a ser grande.

A Dios, quien iluminó mi camino y me dio la sabiduría para culminar esta etapa.

A mis sobrinos, Lina López y Juan Díaz, por darme ánimos para continuar, por ayudarme a crecer, a mi familia, por estar pendiente de cada avance y cada logro, por animarme a seguir.

A mi compañera de vida Jessica, por ayudarme a avanzar, por su apoyo incondicional.

A mi compañeros y amigos, por ayudarme a aprender, por estar siempre en el proceso.

-Samuel Cely

A Dios, por guiar mis pasos, darme la fortaleza para no rendirme y brindarme la sabiduría necesaria para culminar esta etapa de mi vida. A mis padres, Emilce Jaimes y Carlos Moreno, por ser mi pilar fundamental, gracias por su esfuerzo incondicional, por creer en mí desde el primer momento y por brindarme todo el apoyo necesario para llegar hasta aquí. Este logro es tan mío como suyo. A mis abuelos, por su amor infinito, sus consejos y su ejemplo. A mi hermana, Ana Lucía que me inspira profundamente cada día a ser un ejemplo a seguir para ella, a mis amigos, quienes me ayudaron en cada paso de este camino, brindándome su ánimo en los momentos difíciles y compartiendo conmigo las alegrías de este proceso. A todas las personas que formaron parte de mi vida durante esta etapa, aportando a mi aprendizaje y dejando una huella invaluable en mi recorrido dentro de la universidad y mi vida.

-Carlos Moreno

Agradecimientos

A nuestro director, Johann Farith Petit Suárez, por su confianza, por compartirnos sus ideas para el desarrollo de este trabajo, por su ayuda para consolidar este proyecto, a nuestro codirector, Oscar Arnulfo Quiroga Quiroga, por encaminar nuestros avances, por sus consejos y su retroalimentación, por hacer posible este proyecto, al grupo GISEL por el planteamiento de esta idea, por su ayuda en la investigación de este trabajo, al semillero NRSE por permitirnos desarrollar nuestro trabajo y mostrar nuestros avances, por hacer crecer nuestro trabajo, a la Escuela de Ingeniería Eléctrica, Electrónica y de Telecomunicaciones (E3T) y nuestros profesores, por brindarnos conocimiento, por su enseñanza y su apoyo a nuestro crecer profesional a lo largo de esta carrera, A la Universidad Industrial de Santander, por brindarnos nuestra educación, los espacios de aprendizaje y desarrollo personal y profesional, por ser nuestra alma mater.

Tabla de Contenido

Pág.

Lista de Tablas

Pág.

Lista de Figuras

Pág.

Lista de Apéndices

Los apéndices están disponibles en el Repositorio Institucional

Resumen

Título: Simulación del Consumo Energético de Modelos de Inteligencia Artificial en el Marco de la Transición Energética - SIMIAE

Autor: Samuel David Cely Quesada, Carlos Andrés Moreno Jaimes [**]

Palabras Clave: Inteligencia Artificial, Consumo Energético, Sostenibilidad, Emisiones de CO_2, Transición Energética, Green AI, Python.

Descripción: El presente trabajo de grado desarrolla SIMIAE, una herramienta de simulación analítica programada en Python. El objetivo central es cuantificar el consumo de energía eléctrica, las emisiones de dióxido de carbono equivalente (CO_2e) y la huella hídrica asociadas del entrenamiento y la inferencia de modelos de aprendizaje profundo. La metodología integra un enfoque glocal que vincula parámetros técnicos de hardware, como Unidades de Procesamiento Gráfico (GPU) de alta densidad, con las particularidades del Sistema Interconectado Nacional (SIN) de Colombia incluyendo factores de infraestructura (PUE, WUE) para varias regiones.

Los resultados obtenidos mediante los casos de estudio muestran que, para clústeres de GPUs dedicados a tareas de inferencia de gran escala, el consumo anual puede acercarse al orden del gigavatios-hora, con huellas de carbono de cientos de toneladas de CO_2e y costos operativos eléctricos altos, mientras que el volumen de agua requerido para refrigeración puede superar ampliamente el millón de litros, evidenciando que la matriz hidroeléctrica colombiana atenúa las emisiones, pero no el estrés hídrico. Adicionalmente, el benchmarking frente a estudios de referencia de modelos como GPT-3 y Bloom confirma que SIMIAE reproduce con errores inferiores al 10% las estimaciones energéticas reportadas en la literatura, lo que respalda su validez cuantitativa. El proyecto constituye un aporte científico para la sostenibilidad digital, al ofrecer un instrumento transparente y reproducible que permite evaluar escenarios de despliegue de IA y apoyar decisiones de planificación energética y ambiental hasta el año 2035.

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería Eléctrica, Electrónica y Telecomunicaciones. Director: Johann Farith Petit Suárez. Doctor en Ingeniería Eléctrica Electrónica y automática. Codirector: Oscar Arnulfo Quiroga Quiroga. Doctor en Tecnología.

Abstract

Title: Simulation of Energy Consumption of Artificial Intelligence Models within the Framework of the Energy Transition - SIMIAE*

Author: Samuel David Cely Quesada, Carlos Andrés Moreno Jaimes**

Key Words: Artificial Intelligence, Energy Consumption, Sustainability, CO_2 Emissions, Energy Transition, Green AI, Python.

Description: This degree work presents SIMIAE, an analytical simulation tool programmed in Python. The central objective is to quantify the electrical energy consumption, carbon dioxide equivalent (CO_2e) emissions, and associated water footprint of the training and inference of deep learning models. The methodology integrates a glocal approach that links technical hardware parameters, such as high-density Graphics Processing Units (GPUs), with the particularities of Colombia's National Interconnected System (SIN), including infrastructure factors (PUE, WUE) for various regions.

The results obtained through the case studies show that, for GPU clusters dedicated to large-scale inference tasks, annual consumption can approach the gigawatt-hour order of magnitude, with carbon footprints of hundreds of tons of CO_2e and high electrical operating costs, while the volume of water required for cooling can far exceed one million liters, evidencing that Colombia's hydroelectric matrix attenuates emissions, but not water stress. Additionally, benchmarking against reference studies of models such as GPT-3 and Bloom confirms that SIMIAE reproduces the energy estimates reported in the literature with errors below 10%, which supports its quantitative validity. The project constitutes a scientific contribution to digital sustainability, by offering a transparent and reproducible instrument that allows evaluating AI deployment scenarios and supporting energy and environmental planning decisions through the year 2035.

---

* Degree Work

**Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería Eléctrica, Electrónica y de Telecomunicaciones. Director: Johann Farith Petit Suárez, Ph.D. Codirector: Oscar Arnulfo Quiroga Quiroga, Ph.D.

Introducción

La proliferación de modelos de Inteligencia Artificial (IA), en particular las arquitecturas de aprendizaje profundo de gran escala (*Large Language Models* - LLM), ha introducido un nuevo paradigma en el consumo de energía eléctrica a nivel global. El entrenamiento y la inferencia de estos modelos demandan infraestructuras de computación de alto rendimiento (HPC) que operan bajo regímenes de potencia térmica de diseño (TDP) cada vez más exigentes, lo que repercute directamente en la demanda agregada de los centros de datos (de Vries, 2023). A medida que las potencias tecnológicas intensifican esta carrera computacional, la externalidad ambiental derivada de su operación masiva se ha convertido en un desafío crítico para la sostenibilidad digital mundial (Luccioni et al., 2023).

La cuantificación del consumo energético y de las emisiones de dióxido de carbono equivalente (CO_2e) requiere un análisis sistémico que involucre no solo la eficiencia del silicio, sino también el factor de eficacia en el uso de la energía (PUE) de la infraestructura y el factor de emisión de la red eléctrica que la soporta (Patterson et al., 2021). En el contexto colombiano, la transición hacia una matriz energética descarbonizada, delineada en los planes de expansión del Sistema Interconectado Nacional (SIN), ofrece un escenario particular y altamente competitivo frente a los promedios globales de intensidad de carbono (Unidad de Planeación Minero-Energética [UPME], 2020). Sin embargo, la implementación local de estas tecnologías emergentes plantea retos sin precedentes, especialmente en lo referente al estrés hídrico derivado de los sistemas de refrigeración y a la viabilidad económica del mantenimiento de estas infraestructuras. En este sentido, la adopción de principios de *Green AI* resulta imperativa para evaluar el impacto real del desarrollo tecnológico y formular estrategias de mitigación (Schwartz et al., 2020).

Para abordar esta problemática, el presente trabajo de grado tiene como eje central el diseño, desarrollo y validación de SIMIAE, un motor de simulación computacional fundamentado en Python. Esta herramienta permite proyectar matemáticamente el impacto energético, la huella hídrica y el Análisis de Ciclo de Vida (LCA) de las arquitecturas de IA, integrando las variables termodinámicas del hardware de vanguardia con los factores geopolíticos específicos de la red eléctrica en la que operan. El propósito fundamental es proporcionar un instrumento de auditoría técnica que apoye la planificación estratégica y la toma de decisiones en el sector eléctrico y tecnológico.

A través de la aplicación de la herramienta en casos de estudio, validados mediante metodologías de *benchmarking* contra la literatura científica de referencia, los resultados demostraron la robustez y precisión analítica del software. Se evidenció que, si bien la matriz eléctrica renovable de Colombia amortigua significativamente las emisiones operativas frente a clústeres ubicados en regiones fósiles, la dependencia de sistemas de refrigeración evaporativa forzada en latitudes tropicales impone un costo hídrico insostenible. Asimismo, el análisis comprobó que la transición hacia algoritmos optimizados resulta más determinante para la mitigación del impacto ambiental que la fuerza bruta del hardware, sentando las bases para un despliegue de Inteligencia Artificial que sea responsable con los compromisos climáticos del país hacia el año 2035.

.

1. Objetivos

1.1 Objetivo General

Desarrollar una herramienta en Python que simule el consumo energético y las emisiones de carbono asociadas al entrenamiento y uso de modelos de inteligencia artificial, utilizando datos de referencia académicos y técnicos.

1.2 Objetivos Específicos

Diseñar e implementar la calculadora interactiva que permita estimar el consumo energético y las emisiones de carbono a partir de parámetros como número de GPUs, horas de entrenamiento y tamaño de dataset.

Integrar fuentes de datos de referencia (académicos, técnicos y APIs climáticas) que permitan contextualizar las estimaciones según la matriz energética regional y asegurar trazabilidad de los supuestos (datos hasta 2024).

Desarrollar un módulo que compile y armonice los datos disponibles, genere una curva histórica/actual del consumo energético y emisiones asociadas a entrenamiento/uso de IA, y produzca proyecciones anuales hasta el año 2035 bajo al menos tres escenarios (línea base, crecimiento acelerado y mejoras en eficiencia).

Diseñar una interfaz gráfica sencilla en Python que facilite el uso de la herramienta en contextos académicos y de investigación, incluyendo visualizaciones de la curva histórica y las proyecciones hasta 2035.

Validar el prototipo mediante la comparación de sus estimaciones con benchmarks y datos de referencia de la literatura, y documentar las fuentes, supuestos y limitaciones de las proyecciones.

2. Fundamentos Teóricos y Modelado Matemático

2.1 Sistemas de Inteligencia Artificial

La Inteligencia Artificial (IA) se define como la rama de las ciencias computacionales dedicada a diseñar sistemas capaces de emular capacidades cognitivas, tales como el reconocimiento de patrones, el procesamiento del lenguaje natural y la toma de decisiones autónomas (LeCun et al., 2015). Para comprender su impacto energético contemporáneo, es necesario entender su origen y su transición desde la lógica pura hasta el cálculo matricial de alta intensidad.

2.2 Origen y cambio de paradigma

Las bases teóricas de la IA parten de abstracciones matemáticas como el "Perceptrón" de 1958 (Rosenblatt, 1958). Tras el límite alcanzado por los "sistemas expertos" basados en reglas estáticas, la industria migró hacia el Aprendizaje Profundo (*Deep Learning*). En este paradigma, en lugar de programar instrucciones explícitas para resolver problemas complejos, se diseñan Redes Neuronales Artificiales (ANN) que procesan volúmenes masivos de datos para aprender y encontrar correlaciones de forma autónoma (Goodfellow et al., 2016; LeFun et al., 2015).

2.3 Funcionamiento matemático: Red neuronal y Aprendizaje

Matemáticamente, una red neuronal opera como un aproximador universal de funciones no lineales. Su núcleo es la neurona artificial, la cual recibe un vector de datos de entrada, los pondera y aplica una transformación para emitir un resultado:

$$y = \sigma \cdot \sum_{i=1}^{n} (w_i x_i + b)$$

En esta ecuación, x_i representa la información de entrada, w_i son los pesos sinápticos, b es el sesgo y σ es una función de activación no lineal (Goodfellow et al., 2016). El comportamiento de la IA se divide en dos fases críticas que explican su consumo eléctrico:

- **Entrenamiento:** Es el proceso donde la red "aprende". Inicialmente, los pesos ($\boldsymbol{w_i}$) son aleatorios, por lo que la red tiende a cometer errores. Mediante algoritmos de cálculo diferencial (como la retropropagación de errores), el sistema ajusta milimétricamente estos pesos millones de veces hasta minimizar el error frente a los datos esperados (LeCun et al., 2015).
- **Inferencia:** Una vez que la red encuentra los pesos matemáticos correctos (modelo entrenado), se despliega para recibir datos nuevos y predecir resultados, como responder a un usuario en una interfaz determinada.

2.4 La justificación del problema (La Paradoja Energética)

El avance de la IA evidenció una ley de escalabilidad empírica: a mayor cantidad de parámetros y datos, mayor precisión del modelo (Kaplan et al., 2020). Esto provocó que arquitecturas como los Modelos de Lenguaje de Gran Escala (LLM) crecieran hasta poseer cientos de miles de millones de parámetros (Vaswani et al., 2017). Ejecutar matemáticamente estos modelos requiere infraestructuras de supercomputación formadas por GPUs operando al límite de su capacidad térmica (Luccioni et al., 2023). La IA pasó de ser un código abstracto a un proceso termoeléctrico masivo. Para garantizar la digitalización sin comprometer las metas climáticas y la estabilidad del SIN en Colombia, es imperativo desarrollar herramientas como SIMIAE que cuantifiquen y proyecten esta externalidad energética (de Vries, 2023; Durmus Senyapar & Bayindir, 2025).

2.5 Evolución de la Carga Computacional en Modelos de Inteligencia Artificial

La arquitectura de los modelos de aprendizaje profundo (*Deep Learning*), en particular los Modelos de Lenguaje de Gran Escala (LLM) basados en Transformers, ha experimentado un crecimiento exponencial en su tamaño paramétrico. Este fenómeno ha generado una paradoja energética: mientras la IA optimiza las redes eléctricas inteligentes y la predicción de la demanda (Durmus Senyapar & Bayindir, 2025), su propio funcionamiento exige infraestructuras que amenazan con desestabilizar la carga base del sistema eléctrico (Nøland et al., 2024).

El ciclo de vida computacional de un modelo de IA se divide en dos fases críticas, las cuales presentan perfiles de demanda disímiles:

- **Entrenamiento (*Training*):** Es un proceso intensivo y acotado en el tiempo que requiere el procesamiento iterativo de exabytes de datos para el ajuste de pesos sinápticos. Modelos fundacionales contemporáneos, como Llama 3.1 (405B), exigen clústeres de miles de unidades de procesamiento gráfico (GPU) operando a máxima capacidad durante meses, consumiendo decenas de gigavatios-hora (GWh) en un solo ciclo (Meta AI Research, 2024).
- **Inferencia (*Inference*):** Corresponde a la fase operativa donde el modelo entrenado responde a las consultas de los usuarios. Aunque el requerimiento de potencia instantánea por consulta es menor, la inferencia es un proceso continuo y escalable. A largo plazo, el consumo energético agregado de la inferencia supera ampliamente al del entrenamiento, cuantificándose típicamente en kilovatios-hora por millón de *tokens* procesados (Luccioni et al., 2023).

2.6 Infraestructura de Hardware y Potencia Térmica de Diseño (TDP)

El núcleo del consumo eléctrico en los centros de datos recae en el silicio. La métrica fundamental para modelar este consumo es la Potencia Térmica de Diseño (*Thermal Design Power* - TDP), que define la cantidad máxima de calor generado por un componente que el sistema de refrigeración debe disipar bajo cargas de trabajo típicas (de Vries, 2023).

La evolución del hardware orientado a IA ha incrementado dramáticamente el TDP. Históricamente, las arquitecturas como la Nvidia V100 presentaban un TDP cercano a los 300 W. En la actualidad, aceleradores de frontera como la Nvidia B200 (*Blackwell*) alcanzan los 1000 W, y la AMD Instinct MI300X llega a los 750 W. Sin embargo, el modelado preciso no puede limitarse a la multiplicación lineal del TDP de la GPU. En configuraciones de supercomputación, debe incorporarse un factor de penalización por escala u overhead sistémico (O_{sys}). Este factor, que oscila entre 1.30 y 2.10, contabiliza el consumo de los switches de red de alta velocidad (como *InfiniBand*), las unidades de procesamiento de datos (DPU) y los procesadores centrales (CPU) de gestión del nodo (de Vries, 2023; Nøland et al., 2024).

La energía basal demandada por el hardware (E_{hw}) se formula como:

$$E_{hw} = \int_0^t \sum_{i=1}^{n} (P_{GPUi}(\tau) \cdot U_i(\tau) \cdot O_{sys})\, d\tau$$

Donde n es el número de aceleradores, P_{GPU} es la potencia nominal y U es el factor de utilización computacional en el tiempo τ.

2.7 Eficiencia de Infraestructura: El Nexo Agua-Energía (PUE y WUE)

El consumo de los equipos informáticos es solo una fracción del requerimiento total del centro de datos. Para mantener la integridad térmica de los servidores, se requiere un soporte de refrigeración y suministro eléctrico cuya eficiencia se mide mediante el factor de Eficacia en el

Uso de la Energía (*Power Usage Effectiveness* - PUE). El PUE es la relación entre la energía total de la instalación (TDI) y la energía entregada exclusivamente al equipo informático (IT):

$$PUE = \frac{Energía_{TDI}}{Energía_{IT}}$$

Un PUE de 1.0 indica una eficiencia perfecta (teórica). Los centros de datos globales presentan un PUE promedio histórico de 1.67, aunque instalaciones modernas optimizadas para IA alcanzan valores cercanos a 1.15 (Patterson et al., 2021). El consumo simulado final (E_{sim}) en la herramienta SIMIAE se calcula integrando este factor:

$$E_{sim} = E_{hw} \cdot PUE$$

En paralelo, los principios de *Green AI* exigen la cuantificación del estrés hídrico asociado a la refrigeración (Schwartz et al., 2020). Esto se modela mediante el *Water Usage Effectiveness* (WUE), medido en litros de agua evaporada por kilovatio-hora (L/kWh). El WUE es altamente dependiente del clima geográfico: en latitudes frías como Islandia o Noruega, el uso de *Free Cooling* (refrigeración por aire exterior) reduce el WUE a valores cercanos a cero. Por el contrario, en climas cálidos y tropicales, la dependencia de torres de enfriamiento evaporativo puede elevar el WUE a rangos de 1.8 a 2.2 L/kWh (Dong et al., 2025).

Tabla 1

Clasificación de Eficiencia y Estándares de PUE en Centros de Datos

Nivel de Eficiencia	Rango de PUE	Características de la Infraestructura
Alta Eficiencia	1.05 – 1.20	Refrigeración líquida o inmersiva; uso de *Free Cooling*.
Estándar Moderno	1.21 – 1.50	Contención de pasillos fríos/calientes; sistemas HVAC optimizados.
Subóptimo / Heredado	> 1.51	Refrigeración por aire tradicional (CRAC); infraestructura antigua.

Nota. Valores adaptados de los promedios globales de la industria para centros de datos de alta densidad (Patterson et al., 2021; Uptime Institute, 2023).

2.8 Cuantificación de la Huella de Carbono y Análisis de Ciclo de Vida (LCA)

Para evaluar el impacto ambiental real, SIMIAE adopta la metodología del *Greenhouse Gas Protocol* (GHG), diferenciando las emisiones en dos alcances principales dentro del Análisis de Ciclo de Vida (LCA) (Luccioni et al., 2023; Patterson et al., 2021):

Alcance 3 (Carbono Embebido - Manufactura): Emisiones generadas en la manufactura y transporte del hardware. Es un costo de carbono hundido que se amortiza durante la vida útil del equipo.

Alcance 2 (Intensidad de Carbono Operativa): Refleja las emisiones derivadas de la generación de la energía eléctrica consumida. Se modela utilizando el factor de intensidad de carbono de la red eléctrica regional (I_{grid}), expresado en gramos de CO_2 equivalente por kilovatio-hora (gCO_2e/kWh).

Las emisiones operativas totales (CO_2e_{op}) se definen matemáticamente incorporando una tasa de descarbonización anual (R_{decarb}), la cual proyecta el efecto de las políticas climáticas de cada región en su matriz energética a lo largo de los años (t):

$$CO_2e_{op} = E_{sim} \cdot (I_{grid} \cdot (1 - R_{decarb})^t)$$

2.9 Transición Energética y el Sistema Interconectado Nacional (SIN) colombiano

El modelado glocal exige contextualizar los algoritmos de IA dentro de las redes eléctricas específicas donde se despliegan. Mientras las estimaciones globales utilizan promedios de la Agencia Internacional de Energía (IEA) fuertemente influenciados por el uso de carbón y gas natural, el Sistema Interconectado Nacional (SIN) de Colombia presenta una dinámica distinta (UPME, 2020).

La matriz de generación del SIN es predominantemente hidroeléctrica, lo que confiere al país un factor de emisión base relativamente bajo frente a los estándares de Estados Unidos o Asia (Tan et al., 2024; XM S.A. E.S.P., 2024). Sin embargo, la vulnerabilidad climática derivada del fenómeno de El Niño obliga al despacho de plantas termoeléctricas de respaldo, alterando transitoriamente la intensidad de carbono operativa. Para los horizontes de simulación hacia 2035, SIMIAE integra las proyecciones del Plan Energético Nacional de la UPME, asumiendo una tasa de descarbonización conservadora del 2% anual en el SIN, sustentada en la entrada paulatina de Fuentes No Convencionales de Energía Renovable (FNCER), como la solar fotovoltaica y la eólica (UPME, 2020). Esto sitúa a Colombia en un punto de análisis estratégico para el *nearshoring* sostenible de cargas de trabajo computacional, balanceando costos de energía y mitigación de emisiones.

2.10 Modelado Matemático Computacional del Simulador SIMIAE

Para garantizar el rigor cuantitativo exigido en la ingeniería eléctrica, el motor de cálculo de SIMIAE abandona las aproximaciones de consumo estático y adopta un enfoque sistémico multivariable. Este modelo discrimina el origen de la carga computacional, las características de la infraestructura y las variables geopolíticas, estructurando el Análisis de Ciclo de Vida (LCA) en las siguientes dimensiones:

- **Cuantificación de Energía en Infraestructura Local de Hardware:**

Cuando la simulación evalúa un despliegue de clústeres físicos (*Hardware Local*), el consumo total en kilovatios-hora (E_{local}) se modela a partir de la Potencia Térmica de Diseño nominal (P_{nom}) del acelerador (por ejemplo, 1000 W para una Nvidia B200 o 750 W para una AMD MI300X) (de Vries, 2023). Sin embargo, la formulación matemática implementada

incorpora tres factores de ajuste realistas: la cantidad de aceleradores (N_{GPU}), el *overhead* de la topología de red (O_{sys}) y el ciclo de trabajo de la actividad (D_{uty}), expresándose como:

$$E_{local} = \left(\frac{P_{nom} \cdot N_{GPU} \cdot O_{sys}}{1000}\right) \cdot t \cdot PUE \cdot D_{uty}$$

El factor O_{sys} escala de forma no lineal. En una estación de trabajo básica, el consumo auxiliar (CPU, RAM, ventilación local) representa un recargo del 30% (O_{sys} = 1.30). En contraste, en arquitecturas de supercomputación (*Super-Clúster* de 1024 GPUs), la inclusión de *switches* InfiniBand y sistemas de almacenamiento masivo eleva la penalización a un 110% adicional (O_{sys} = 2.10) (Luccioni et al., 2023; Uptime Institute, 2023). Por su parte, la variable D_{uty} diferencia la exigencia de las tareas: durante el entrenamiento asume un valor continuo de 1.0 (carga térmica máxima sostenida), mientras que en tareas de inferencia se establece en 0.4, reflejando los estados de latencia entre peticiones (Dodge et al., 2022).

- **Estimación Basada en Modelos Fundacionales (LLM):**

Para escenarios donde no se especifica el hardware físico sino el uso de arquitecturas de lenguaje entrenadas previamente (*Modelo Externo*), el simulador fundamenta el cálculo en datos estandarizados de la industria (Meta AI Research, 2024). La energía consumida (E_{LLM}) se determina mediante un factor de demanda específico del modelo (C_{base}), multiplicado por el volumen de datos procesados (D_{val}) y la eficiencia del centro de datos (PUE):

$$E_{LLM} = D_{val} \cdot C_{base} \cdot PUE$$

En la fase de entrenamiento, D_{val} representa los ciclos completos y C_{base} se mide en kilovatios-hora masivos (e.g., 22.2 GWh para Llama 3 405B). En inferencia, D_{val} equivale a los millones de *tokens* generados, y C_{base} oscila entre 0.35 kWh para arquitecturas eficientes (como Gemini 1.5 Pro) y 3.96 kWh para modelos densos como Bloom (Brown et al., 2020).

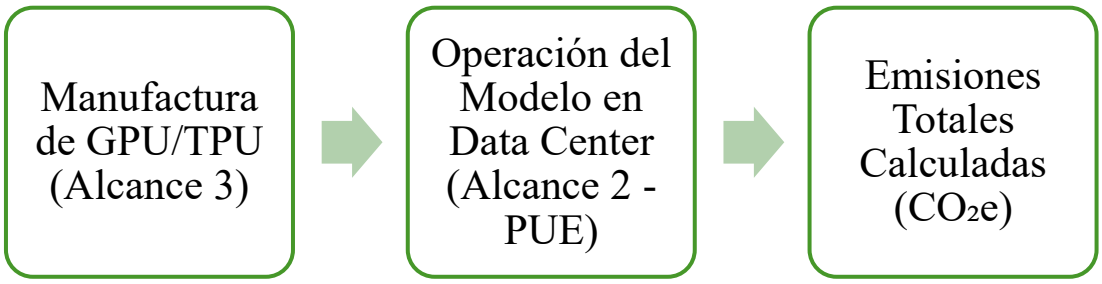
- **Análisis de Emisiones LCA**

El impacto ambiental se disgrega en dos alcances normativos fundamentales. Las emisiones operativas (Alcance 2), derivadas del consumo eléctrico de la red, se determinan cruzando la energía simulada con el factor de emisión local (I_{grid}):

$$CO2_{op} = \frac{E \cdot I_{grid}}{1000}$$

Figura 1

Fronteras del Análisis de Ciclo de Vida (LCA) modeladas en SIMIAE



Nota: El diagrama ilustra los límites del sistema adoptados por el motor de cálculo.

Para cumplir con los protocolos avanzados de sostenibilidad (LCA), el programa cuantifica el carbono embebido o de manufactura (Alcance 3). Este cálculo se aplica exclusivamente a escenarios de hardware local y depende de un factor de fabricación fijo por arquitectura (C_{emb}), que varía desde 90 $kgCO_2$ para GPUs comerciales hasta 280 $kgCO_2$ para aceleradores de centro de datos (Luccioni et al., 2023):

$$CO2_{emb} = C_{emb} \cdot N_{GPU}$$

El impacto climático total del ciclo de simulación corresponde a la suma de ambas componentes operativas y de fabricación.

- **Modelado Dinámico de la relación Agua-Energía (WUE):**

Una contribución significativa del simulador es la regionalización de la huella hídrica. A diferencia de las estimaciones estáticas, SIMIAE incluye un factor binario de tecnología de refrigeración (F_{tech}). Si el centro de datos opera en latitudes que permiten *Free Cooling* (como Islandia o Noruega, con un PUE < 1.15), el simulador asume que la evaporación forzada de agua es nula (F_{tech} = 0). Para el resto de las regiones, incluyendo el SIN colombiano, se aplica el factor hídrico base (WUE_{base}) para calcular el consumo total de agua (W) (Google, 2024):

$$W = E \cdot WUE_{base} \cdot F_{tech}$$

- **Dinámica Temporal: Retroproyección y Transición Energética al 2035:**

Para evaluar el impacto prospectivo y la evolución tecnológica, el núcleo analítico de SIMIAE incorpora dos familias de proyecciones matemáticas temporales (donde *t* representa el diferencial de años respecto a la línea base de 2024).

La primera familia de proyecciones corresponde a las curvas de demanda energética ($\boldsymbol{E_{futuro}}$). Se establecen tres escenarios de crecimiento para la demanda base (E_0). Un escenario referencial (r = 0.15), un escenario *Hyper AI* de adopción acelerada (r = 0.30) y un escenario *Green AI* sustentado en optimizaciones algorítmicas como la cuantización (r = 0.02) (Schwartz et al., 2020):

$$E_{futuro}(t) = E_0 \cdot (1 + r)^t$$

La segunda familia aborda la descarbonización de la red eléctrica. Las proyecciones de emisiones integran obligatoriamente las políticas de transición de cada país. Se aplica una tasa de descarbonización anual (R_{decarb}) que reduce progresivamente la intensidad de carbono del kilovatio-hora, simulando la entrada de Fuentes No Convencionales de Energía Renovable (FNCER) al sistema (Durmus Senyapar & Bayindir, 2025; UPME, 2020):

$$I_{grid}(t) = I_{grid0} \cdot (1 - R_{decarb})^t$$

3. Diseño Metodológico y Computacional

El presente trabajo de grado se desarrolló bajo un enfoque metodológico cuantitativo de simulación analítica e ingeniería de software científico. La investigación se estructuró como un modelo sistémico híbrido, combinando la fundamentación teórica de los sistemas de potencia y el análisis de ciclo de vida (LCA) ambiental, con el diseño de una arquitectura computacional interactiva. El objetivo metodológico consistió en traducir la fenomenología termodinámica de los centros de datos en un motor de cálculo matemático codificado en Python.

A continuación, se detallan los componentes metodológicos que rigen el funcionamiento del simulador SIMIAE, estructurados en participantes (variables del sistema), herramientas (ecosistema de software) y procedimientos (arquitectura algorítmica).

3.1. Participantes (Entidades Computacionales y Geopolíticas)

Esta sección define los participantes del modelo matemático de SIMIAE, estructurados en tres dimensiones operativas: física (hardware), lógica (modelos de IA) y geopolítica (redes eléctricas). Su parametrización en bases de datos internas garantiza que las proyecciones operen con especificaciones técnicas e indicadores ambientales reales.

3.1.1. Hardware de Cómputo de Alto Rendimiento (Base de Datos de Potencia)

La capa física de la simulación está constituida por una selección representativa de Unidades de Procesamiento Gráfico (GPU) y Unidades de Procesamiento Tensorial (TPU) que definen el estado del arte en aceleración de Inteligencia Artificial. La métrica rectora para el modelado de demanda máxima es la Potencia Térmica de Diseño (TDP), la cual indica el límite superior de disipación de calor bajo cargas de trabajo sostenidas en operaciones matriciales de precisión mixta (FP8, FP16, BF16).

La parametrización integrada en el diccionario *Hardware_Db* de la herramienta extrae sus valores fundacionales directamente de las hojas de datos (*datasheets*) oficiales de los fabricantes de semiconductores, garantizando la fidelidad industrial del modelo:

- **Arquitectura Nvidia Blackwell (B200):** Representa el límite superior de consumo actual. Parametrizada con un TDP de 1000 W, esta arquitectura integra un empacado 2.5D avanzado que enlaza múltiples matrices de silicio, exigiendo refrigeración líquida directa al chip (D2C) en configuraciones de clúster (Nvidia, 2024).
- **Arquitectura AMD CDNA 3 (Instinct MI300X):** Integrada al simulador con un TDP nominal de 750 W. Se caracteriza por su alta densidad de memoria HBM3, lo que desplaza gran parte del consumo energético hacia los controladores de memoria de alto ancho de banda durante la inferencia de modelos masivos (Advanced Micro Devices, 2023).
- **Arquitectura Nvidia Hopper (H100 - SXM5):** Parametrizada con un TDP de 700 W, sirviendo como el estándar industrial primario para la base comparativa de la simulación frente a hardware de generaciones previas (Nvidia, 2022).
- **Equipos Legados y Comerciales:** Se incluyen modelos como la Nvidia A100 (400 W), V100 (300 W) y hardware de nivel consumidor (*Consumer-grade*) como la RTX 4090 (450 W) para simular escenarios de *Fine-Tuning* (ajuste fino) en estaciones de trabajo universitarias o laboratorios locales.

De manera paralela, el simulador asigna a cada uno de estos participantes una huella de carbono embebida (Alcance 3) en el diccionario *Embodied_Carbon_Db*, escalonada desde 90 $kgCO_2$ hasta 280 $kgCO_2$ por chip, reflejando la intensidad energética de la fotolitografía extrema ultravioleta (EUV) requerida para la fabricación de nodos de 4 nm y 3 nm (Luccioni et al., 2023).

3.1.2. Modelos Fundacionales de Inteligencia Artificial (LLM)

La capa algorítmica de los participantes está definida por los Modelos de Lenguaje de Gran Escala (LLM). Estos representan la "carga de trabajo" o demanda lógica que estresa el hardware. Para evitar generalizaciones imprecisas, el simulador distingue drásticamente el consumo entre la fase de entrenamiento (medida en gigavatios-hora por ciclo) y la fase de inferencia (medida en kilovatios-hora por millón de *tokens*).

El diccionario *Model_Zoo* parametriza arquitecturas que cubren diferentes paradigmas de eficiencia computacional:

- **Entrenamiento de Frontera (Modelos Densos):** El participante principal de calibración masiva es **Llama 3.1 (405B)** de Meta. Este modelo denso requirió un clúster de más de 16,000 GPUs operando simultáneamente durante meses, resultando en un consumo parametrizado en la herramienta de 22.2 GWh por ciclo de entrenamiento completo, una métrica extraída del reporte técnico oficial de la arquitectura (Meta AI Research, 2024).
- **Inferencia y Eficiencia Arquitectónica:** Durante la fase operativa, el tamaño de los parámetros dicta la intensidad energética. Modelos densos y tempranos como **Bloom (176B)** o **GPT-3 (175B)** presentan factores de inferencia altos, superando los 3.0 kWh por millón de *tokens* generados debido a que activan la totalidad de su red en cada consulta (Brown et al., 2020; Luccioni et al., 2023). En contraste, se introducen modelos más modernos en el simulador (como **Gemini 1.5 Pro** o **GPT-4o**) que, al implementar arquitecturas de Mezcla de Expertos (MoE - *Mixture of Experts*), enrutan la computación activando solo una fracción de sus parámetros por *token*, logrando eficiencias de inferencia parametrizadas hasta en 0.35 kWh por millón de *tokens*.

3.1.3. Escenarios Geopolíticos y Redes Eléctricas (El Enfoque Glocal)

El diccionario *Grid_Data* actúa como el marco glocal de la simulación. Para el Sistema Interconectado Nacional (SIN) de Colombia, se asignó una intensidad de carbono de 164 gCO_2e/kWh y una tasa de descarbonización (*decarb_rate*) del 2% anual, alineada con la UPME y XM S.A. (UPME, 2020; XM S.A. E.S.P., 2024). Se contrastó con redes de Estados Unidos (Texas, Virginia) validadas por la EPA, y de Europa (Noruega, Islandia) donde prevalece la hidroeléctrica y geotérmica, permitiendo factores PUE ultrabajos y el uso de *Free Cooling*.

3.1.4. Penalización por Escala e Infraestructura (Overhead)

En el diccionario *Infra_Scale*, se implementó un factor de ajuste que varía desde 1.30 para estaciones de trabajo (*Workstations*) hasta 2.10 para infraestructuras *Super-Clúster* de 1024 GPUs, contabilizando el consumo de la topología de red *InfiniBand* y la refrigeración del bastidor (*rack*) (de Vries, 2023; Uptime Institute, 2023).

3.2. Herramientas (Ecosistema de Software y Arquitectura Computacional)

El desarrollo del simulador SIMIAE se estructuró sobre el lenguaje de programación Python, seleccionado por su preeminencia en la computación científica, su acceso simplificado y su capacidad para integrar procesamiento matricial con interfaces web interactivas. La herramienta se divide en dos módulos interdependientes: el núcleo analítico (motor.py) y el microservicio de interfaz gráfica (app_ui.py). A continuación, se detalla el ecosistema de dependencias y la implementación algorítmica de la aplicación.

3.2.1. Ecosistema de Librerías de Python Utilizadas

Para garantizar la eficiencia computacional y la calidad gráfica, el desarrollo importó un *stack* tecnológico especializado, cuyas librerías se justifican académicamente de la siguiente manera:

- **NumPy (numpy):** Fundamental para el motor matemático. Permitió vectorizar las proyecciones temporales de consumo y emisiones, reemplazando bucles escalares ineficientes por operaciones de álgebra lineal sobre *arrays* a través de funciones como *np.arange*, *np.cumsum* y *np.interp* (Harris et al., 2020).
- **Pandas (pandas):** Utilizado para la estructuración analítica de datos (*DataFrames*). Su función principal en SIMIAE es la manipulación de las series temporales interpoladas y la preparación de matrices de datos tabulares para su posterior exportación a formatos de auditoría (McKinney, 2010).
- **Dash y extensiones (*dash, dcc, html*):** *Framework* principal de la arquitectura *Front-End*. Permitió construir una aplicación web analítica de una sola página (SPA) sin recurrir a lenguajes externos como JavaScript, gestionando la reactividad mediante el enrutamiento de estados (*Input, Output, State*).
- **Plotly Graph Objects (*plotly.graph_objects*):** Motor de renderizado vectorial interactivo (basado en WebGL y D3.js). Se implementó para la creación de cartografía térmica dinámica (mapas coropléticos) y curvas de proyección prospectivas, permitiendo interactividad nativa (*hover*, *zoom*, aislamientos de trazas).
- **Matplotlib (*matplotlib.pyplot*):** En contraste con Plotly (orientado a web), Matplotlib se configuró en modo *backend* (usando *matplotlib.use('Agg')*) para renderizar gráficas estáticas de alta resolución gráfica en formato JPG, diseñadas exclusivamente para ser incrustadas en los reportes físicos documentales (Hunter, 2007).
- **FPDF y OpenPyXL (*fpdf*, *openpyxl*):** Librerías de gestión documental. *FPDF* asume la vectorización y el diseño de la planimetría del Dossier Ejecutivo en formato PDF, mientras

que *OpenPyXL* permite acceder a nivel de celda en Excel para inyectar diseño condicional (colores corporativos, fuentes, formatos numéricos).

3.2.2. Núcleo Matemático y Algorítmico (motor.py)

El módulo motor.py consolida la lógica de negocio y las ecuaciones termodinámicas del proyecto. Su implementación se estructura en tres procedimientos principales:

- **Cuantificación Base Multivariable (*calculate_base_metrics*):** Esta función procesa el vector de parámetros del usuario. Mediante un flujo condicional, evalúa si la métrica de entrada corresponde a *Hardware Local* (cálculo integrando TDP temporal, escalamiento de infraestructura y ciclo de trabajo) o a Modelo Externo *(LLM)* (cálculo basado en coeficientes de consumo por ciclo masivo o inferencia de *tokens*). Seguidamente, aplica los multiplicadores de *Power Usage Effectiveness* (PUE), extrae la intensidad de carbono georreferenciada (*grid['co2']*) e implementa la interpretación del nexo agua-energía, anulando el impacto hídrico (*factor_tecnologia = 0.0*) cuando el PUE indica refrigeración natural (*Free Cooling* en latitudes nórdicas).
- **Sistema de Prospección Exponencial (*get_projections* y *get_historical_curve*):** Utilizando el arreglo temporal base de *NumPy*, el algoritmo proyecta el horizonte hacia 2035 simulando tres escenarios de carga: Línea base (crecimiento del 15% anual), *Hyper AI* (30% anual) y *Green AI* (2% anual condicionado por optimización algorítmica). En paralelo, aplica la variable *decarb_rate* para decrementar anualmente el factor de emisión. Para el análisis histórico (2014-2024), el código diferencia la Ley de Koomey (hardware) mediante curvas logarítmicas (*np.logspace*), frente a la explosión exponencial inversa de los parámetros LLM.

- **Optimizador Inteligente Heurístico (*ejecutar_optimizador_inteligente*):** Algoritmo de búsqueda que itera sobre la base de datos geopolítica (*Grid_Data*). Tras filtrar restricciones normativas (p. ej., soberanía de datos que bloquea nodos en EE.UU. o China), simula internamente todas las combinaciones geográficas posibles para retornar la ruta de mínimo impacto de carbono (*Verde*), mínimo costo operativo (*Económica*) y balance normativo (*Segura*).

3.2.3. Arquitectura de Interfaz y Reactividad (app_ui.py)

El comportamiento de la interfaz gráfica no es procedimental, sino reactivo basado en el patrón de diseño Observador. El código gestiona el flujo de datos a través del cerebro principal de *callbacks* (*ejecutar_todo*).

- **Inyección de Dependencias y Estados:** La función recolecta el estado actual de hasta 22 variables en pantalla (*State*) disparadas por eventos de usuario (*Input*, como clics o cambios de selectores). El algoritmo discrimina dinámicamente si el "Modo de Análisis" es aislado (*Single*) o comparativo (*Compare*), para instanciar una o dos llamadas simultáneas al código *motor.py*.
- **Renderizado Vectorial (*Plotly*):** Con los resultados devueltos, el sistema instancia objetos *go.Figure()*. Para el componente geoespacial, inyecta los resultados en un *go.Choropleth* utilizando códigos ISO-3166-1 alfa-3 (ej. 'COL' para Colombia, 'USA' para EE.UU.), coloreando las regiones según los escenarios (Verde y Dorado).
- **Interpolación Temporal:** Para suavizar la transición visual de las proyecciones en las gráficas de línea, se implementó el método *interpolar_mensual*, el cual utiliza *pd.date_range* para transformar la matriz anual proyectada en una curva mensual continua, procesada algebraicamente por *np.interp* para el trazado detallado.

3.2.4. Automatización de Exportación y Reportes Científicos

Para asegurar la trazabilidad institucional y el control de calidad documental de los análisis generados, *app_ui.py* implementa dos procesos de compilación de archivos:

- **Estructuración Analítica (*export_excel*):** Un *callback* independiente captura el estado de simulación e instancia un canal en memoria (*io.BytesIO()*). Utilizando *pd.ExcelWriter*, el algoritmo escribe múltiples hojas (*sheets*): una de resumen consolidado y varias correspondientes a las proyecciones temporales interpoladas. El motor de *OpenPyXL* interviene secuencialmente (*ws.iter_rows*) para formatear encabezados, aplicar diseño condicional en cebra (*Zebra-striping*), ajustar automáticamente el ancho de las columnas según la dimensión de los datos y codificar las celdas financieras de forma nativa.
- **Generación del Dossier Ejecutivo (*export_pdf*):** La exportación a PDF se ejecuta sobre una clase extendida PDF (*FPDF*) que define encabezados y pies de página estandarizados. A diferencia de las exportaciones dinámicas de *Plotly*, este proceso compila una gráfica estática *ad-hoc*; se instancia un objeto de sub gráficos (*plt.subplots*) de 2x2 en *Matplotlib* para resumir Energía, Emisiones, Costos y Agua. La figura es renderizada fuera de pantalla, exportada temporalmente como *raster* (*temp_simiae_chart.jpg*), incrustada métricamente en las coordenadas del documento FPDF mediante el método *pdf.image*, y finalmente procesada en bytes para su entrega asíncrona al navegador del usuario, eliminando los residuos en el servidor de forma segura (*os.remove*).

3.3. Procedimientos (Arquitectura Algorítmica y Flujo de Ejecución)

El procedimiento metodológico para la construcción del simulador SIMIAE se estructuró en tres fases de ingeniería secuenciales: la consolidación del motor matemático, el desarrollo de los algoritmos de prospección temporal, y la implementación de la arquitectura reactiva de interfaz. Esta metodología garantiza que las ecuaciones de estado termodinámico y ambiental se ejecuten con precisión vectorial.

3.3.1. Fase 1: Estructuración Matemática y Evolución del Motor de Cálculo (motor.py)

El primer procedimiento consistió en la codificación del núcleo analítico del programa. Inicialmente, el proyecto partió de un borrador estático de multiplicaciones lineales. Sin embargo, para cumplir con el rigor del Análisis de Ciclo de Vida (LCA), el algoritmo evolucionó hacia un modelo de direccionamiento condicional encapsulado en la función principal *calculate_base_metrics*.

- **Ingesta y Parametrización de Datos:** El procedimiento inicia cargando en memoria RAM los diccionarios de constantes físicas (*Hardware_Db, Grid_Data, Model_Zoo*). Estos diccionarios actúan como la base de datos no relacional del sistema, almacenando, por ejemplo, los 1000 W de TDP de la arquitectura Blackwell [19] o el factor de emisión de 164 gCO_2e/kWh del SIN colombiano [31]. La precisión del motor analítico depende de la veracidad de los coeficientes técnicos asignados a cada arquitectura de hardware. Para este fin, se consolidó una base de datos interna que asocia el consumo de potencia basal y la intensidad de fabricación de los aceleradores más representativos de la industria. Los valores de Potencia Térmica de Diseño (TDP) y la huella de carbono embebida para los dispositivos integrados en el simulador se detallan en la Tabla 2.

Tabla 2

Especificaciones Técnicas de los Aceleradores gráficos de Hardware

Arquitectura de Hardware	**TDP Nominal (W)**	**Carbono Embebido ($kgCO_2e$)**	**Hoja de datos (Anexo)**
Nvidia B200 (Blackwell 2024)	1000	280.0	(Anexo B) (Nvidia, 2024)
Nvidia H100 (2023)	700	280.0	(Anexo C) (Nvidia, 2022)
AMD MI300X (2024)	750	250.0	(Anexo D) (AMD, 2023)
Google TPU v5p (2023)	450	250.0	(Google Cloud, s.f.)
Nvidia A100 (2020)	400	150.0	(Anexo E) (Nvidia, 2020)
Nvidia V100 (2017)	300	100.0	(Anexo F) (Nvidia, 2017)
RTX 4090 (Consumer)	450	90.0	(Nvidia, s.f.)
RTX 3090 (Consumer)	350	90.0	(Nvidia, s.f.)

Nota: Los valores de TDP corresponden al consumo máximo sostenido por unidad. El carbono embebido representa la estimación de emisiones de Alcance 3 durante la manufactura.

Observación documental: Para garantizar la trazabilidad, reproducibilidad y rigor metodológico del trabajo realizado, las especificaciones técnicas originales (*datasheets*) correspondientes a cada arquitectura de hardware han sido adjuntadas en su totalidad desde el **Anexo B** hasta el **Anexo F** del presente trabajo de grado.

- **Bifurcación del Flujo de Trabajo (Hardware vs. LLM):** El algoritmo evalúa el parámetro de entrada *t_recurso*. Si el análisis es sobre infraestructura física, el flujo deriva hacia la ecuación de hardware local. El código multiplica la potencia nominal del acelerador por la cantidad de GPUs (*scale_gpus*) y aplica el *overhead* sistémico (*scale_overhead*) extraído de la topología seleccionada, integrando el ciclo de trabajo (*duty*) de 1.0 para entrenamiento y 0.4 para inferencia. Por el contrario, si el usuario evalúa un *Modelo Externo*, el flujo se redirige a multiplicar el volumen de datos (*val_data*) por los factores de consumo masivo previamente calibrados del modelo (p. ej., los 22.2 GWh de Llama 3) (Meta AI Research, 2024).
- **Cálculo de Externalidades (PUE, WUE y Emisiones):** Una vez consolidado el consumo energético en kilovatios-hora, el procedimiento inyecta las ineficiencias de la infraestructura. Multiplica la energía por el factor PUE regional. Seguidamente, ejecuta la lógica del nexo agua-energía: evalúa condicionalmente si el PUE es inferior a 1.15; si la condición se cumple, asigna la variable *factor_tecnologia* = 0.0 (simulando *Free Cooling*), anulando el consumo hídrico operativo. Finalmente, disgrega la huella de carbono separando la multiplicación directa del factor de red (Alcance 2) y la extracción del carbono de manufactura desde *Embidied_Carbon_Db* (Alcance 3) (Luccioni et al., 2023).

3.3.2. Fase 2: Modelado de Proyecciones Sistémicas y Optimizador Heurístico

La segunda fase metodológica consistió en agregar al simulador profundidad temporal y capacidad de toma de decisiones, superando la limitación de los cálculos estáticos del código inicial trabajado.

- **Vectorización de Proyecciones Prospectivas:** Mediante la función *get_projections*, se implementó la proyección al año 2035. Para optimizar el rendimiento computacional, se reemplazaron los bucles iterativos convencionales por operaciones vectoriales utilizando la librería *NumPy* (Harris et al., 2020). Se generó un *array* temporal (*np.arange(2024, 2036)*) sobre el cual se aplicaron algebraicamente las tasas de crecimiento. El procedimiento ejecuta en paralelo tres curvas exponenciales de la forma $\boldsymbol{E_0 \cdot (1 + r)^t}$: un escenario base (15% anual), un escenario *Hyper AI* (30%) y un escenario *Green AI* (2%) (Schwartz et al., 2020). Simultáneamente, el vector de intensidad de carbono se atenúa año tras año aplicando la tasa de descarbonización (*decarb_rate*) específica de la red evaluada (UPME, 2020).
- **Dinamismo de la Curva Histórica:** El procedimiento *get_historical_curve* resolvió la dualidad tecnológica hacia el pasado (2014-2024). Para el hardware, se codificó una función logarítmica (*np.logspace*) que simula la Ley de Koomey, demostrando que en el pasado se requería más energía para el mismo cómputo. Para los modelos LLM, se implementó una reducción exponencial inversa (división por potencias de 2), calibrando el algoritmo para reflejar que la carga de entrenamiento masivo era inexistente hace una década (Durmus Senyapar & Bayindir, 2025).
- **Búsqueda Heurística (Optimizador Inteligente):** Se programó un procedimiento de evaluación de escenarios (*ejecutar_optimizador_inteligente*). Este algoritmo itera sobre la totalidad del diccionario geopolítico global. Primero, aplica filtros restrictivos (bloqueando EE.UU. o China si el usuario selecciona privacidad estricta bajo el *Cloud Act*). Luego, calcula silenciosamente las métricas para todos los países válidos y ejecuta una función de ordenamiento iterativo (*sorted*), retornando automáticamente los

clústeres que minimizan el costo financiero (*Ruta Económica*), la latencia o la huella de carbono absoluta (*Ruta Verde*).

3.3.3. Fase 3: Flujo de Interacción, Reactividad y Exportación Documental (app_ui.py)

La fase final del procedimiento consistió en encapsular el motor matemático dentro de una arquitectura web reactiva, permitiendo la exploración de datos sin recargar el servidor.

- **Arquitectura de Microservicios y Callbacks:** El *software* se estructuró sobre *Dash*, utilizando el patrón de diseño Observador. El procedimiento maestro, el decorador *@app.callback* asociado a la función *ejecutar_todo*, actúa como el orquestador del flujo. Se comunica simultáneamente 22 estados de la interfaz (*State*). Al recibir un evento de cálculo (*n_clicks*), invoca al motor, procesa la respuesta (diccionarios *JSON*), y actualiza asíncronamente los contenedores HTML (*KPIs*, textos de asesoría).
- **Renderizado Gráfico e Interpolación:** Para la visualización temporal, los *arrays* anuales provenientes del motor matemático son crudos. El procedimiento implementa la función *interpolar_mensual*, la cual utiliza Pandas (*pd.date_range*) y *NumPy* (*np.interp*) para discretizar la serie en 120 puntos intermedios, suavizando las curvas de *Plotly* para un análisis visual preciso (Harris et al., 2020; McKinney, 2010). El mapa coroplético se renderiza mapeando las selecciones del usuario a códigos ISO de la cartografía interna de *Plotly*.
- **Flujo de Exportación Científica (PDF y Excel):** Para exportar la base de datos, el código invoca un flujo en memoria (*io.BytesIO*), donde *Pandas* escribe múltiples hojas de cálculo. Posteriormente, la librería *OpenPyXL* itera sobre las celdas aplicando formato condicional automatizado (diseño en cebra, formato moneda y ajustes de ancho de columna). Para el *Dossier PDF*, el procedimiento es híbrido: utiliza *Matplotlib* en

modo *backend* para procesar una matriz estática de gráficos de barras 2x2, la exporta temporalmente como imagen procesada de alta definición (Hunter, 2007), la incrusta estratégicamente en el lienzo vectorial generado por *FPDF* y, tras la conversión a bytes de descarga, ejecuta un procedimiento de recolección de basura (*os.remove*) para purgar el archivo temporal del servidor.

Observación documental: El programa SIMIAE.exe y sus códigos base motor.py y app_ui.py creados a partir del análisis previo han sido adjuntados en el **Anexo A** del presente trabajo de grado.

4. Resultados y análisis.

En este capítulo se presentan los resultados obtenidos tras la ejecución de las simulaciones computacionales en la herramienta SIMIAE. La evaluación se fundamenta en la aplicación de casos de estudio representativos que reflejan las condiciones operativas de la infraestructura tecnológica en el marco de la transición energética. El análisis cualitativo y el tratamiento cuantitativo de los datos permiten validar las ecuaciones de estado térmico, el Análisis de Ciclo de Vida (LCA) y las proyecciones de emisiones establecidas en el planteamiento metodológico.

4.1. Caso de Estudio 1: Análisis Anual de Inferencia HPC en el SIN colombiano

El primer caso de estudio tiene como objetivo cuantificar el impacto energético, ambiental y financiero del mantenimiento operativo de una infraestructura de Computación de Alto Rendimiento (HPC) dedicada a tareas de inferencia durante una anualidad completa (8,760 horas). Se configuró un escenario aislado bajo los parámetros técnicos de vanguardia que se detallan en la Tabla 3.

Tabla 3

Parámetros de Configuración del Motor de Cálculo - Caso de Estudio 1

Variable de Simulación	Parámetro Ingresado
Modo de Análisis	Un Escenario Aislado (Single)
Restricción Normativa	Libre Flujo
Actividad Computacional	Inferencia
Modalidad	Hardware Local
Arquitectura de Hardware	Nvidia H100 (2023) - TDP Nominal: 700 W
Escala de Infraestructura	HPC (128 GPUs)
Tiempo de Uso Sostenido	8,760 horas (1 Año)
Ubicación del Data Center	Colombia (SIN - Hidro/Térmica)
Eficiencia de la Instalación	Estándar (PUE = 1.58)
Uso optimización industrial IEA	No

Nota: Datos extraídos del módulo de configuración del simulador SIMIAE. La escala HPC de 128 GPUs incluye un factor de penalización sistémica (*overhead*) de 1.95 para contabilizar la interconexión de red y refrigeración.

4.1.1. Análisis Cuantitativo y Tratamiento de Resultados

Al ejecutar el motor matemático para el periodo de un año, el simulador integró las variables de potencia de la arquitectura Nvidia H100 con los factores geopolíticos del Sistema Interconectado Nacional colombiano (SIN). La Tabla 4 consolida las métricas de impacto resultantes, evidenciando el consumo masivo derivado de la operación ininterrumpida de 128 procesadores gráficos.

Tabla 4

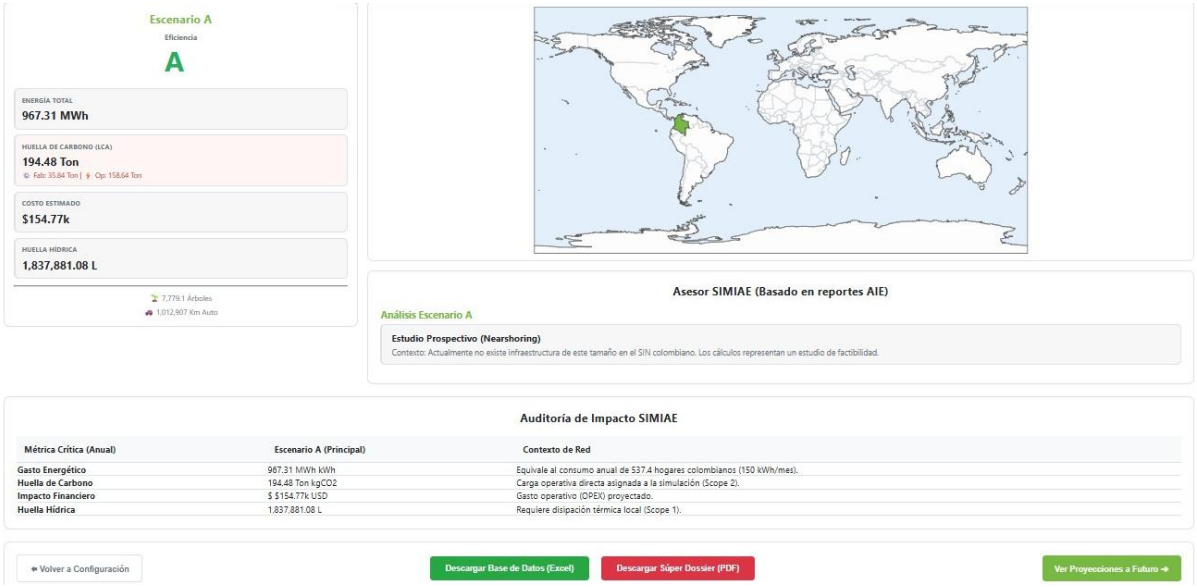
Métricas de Impacto y Análisis de Ciclo de Vida (LCA) - Caso de Estudio 1

Métrica de Impacto	Valor Cuantificado	Unidad de Medida
Calificación de Eficiencia	A (Color Verde)	Escala SIMIAE
Energía Total Consumida	967,305.83	kWh (0.96 GWh)
Huella de Carbono Total (LCA)	194,478.16	$kgCO_2e$
- *Carbono Embebido (Fab)*	35,840.00	$kgCO_2e$ (Alcance 3)
- *Carbono Operativo (Op)*	158,638.16	$kgCO_2e$ (Alcance 2)
Costo Operativo Estimado	154,768.93	USD
Huella Hídrica Total	1,837,881.08	Litros

Nota: Resultados calculados para una intensidad de carbono de 164 gCO_2e/kWh (factor de red Colombia) y una tarifa eléctrica industrial de 0.16 USD/kWh (XM S.A. E.S.P., 2024).

Figura 2

Panel de Resultados y Asesoría Interactiva para el Caso de Estudio 1



Nota: La interfaz gráfica muestra el desglose del Análisis de Ciclo de Vida (LCA) para 128 GPUs Nvidia H100 operando en Colombia. Las equivalencias ambientales indican que mitigar este

impacto requeriría la siembra de 7,779.1 árboles o evitar el recorrido de 1,012,907 km en un vehículo a combustión.

4.1.2. Análisis Cualitativo del Desempeño Energético y Ambiental

El análisis de SIMIAE revela que la duración operativa (8,760 horas) es el principal determinante de la carga eléctrica, generando un consumo industrial masivo de 0.96 GWh. Aunque la baja intensidad de carbono de la matriz colombiana amortigua las emisiones operativas otorgándole una Calificación de Eficiencia 'A', el costo anual de $154,768 USD desafía la viabilidad económica del proyecto.

Paralelamente, el modelo LCA cuantificó el Alcance 3 en 35,840 $kgCO_2$, correspondientes a la fabricación de 128 GPUs Nvidia H100 (Nvidia, 2022). Además, se identificó un estrés hídrico crítico de 1.83 millones de litros; este volumen demuestra que operar en Colombia bajo un PUE de 1.58 depende estrictamente de refrigeración evaporativa, un impacto que solo podría mitigarse transitando hacia tecnologías de refrigeración líquida o climas aptos para *Free Cooling*.

Figura 3

Proyección Dinámica de Consumo Anual y Emisiones - Caso de Estudio 1

Nota: Gráfica vectorial generada que ilustra la divergencia de consumo proyectada hasta el año 2035 para el escenario de inferencia anual.

Observación documental: Para validar la trazabilidad del software desarrollado, los reportes completos de configuración, el Dossier Ejecutivo (PDF) y las matrices de bases de datos de interpolación mensual (Excel/CSV) generados en este análisis han sido adjuntados en el **Anexo A** del presente trabajo de grado.

4.2. Caso de Estudio 2: Validación por Benchmarking Multi-estudio (GPT-3 vs. Bloom)

El segundo caso de estudio tiene como propósito fundamental validar la precisión del motor de cálculo de SIMIAE mediante la contrastación de sus resultados con los dos estudios de referencia más robustos en la literatura científica actual sobre la huella energética de modelos fundacionales. Se seleccionó un escenario de comparación de arquitecturas masivas (LLM) con el fin de verificar la respuesta de las ecuaciones de integración temporal frente a datos reales auditados de entrenamiento. Los parámetros técnicos ingresados para la validación se detallan en la Tabla 5.

Tabla 5

Parámetros Comparativos de Configuración - Caso de Estudio 2

Variable de Simulación	Escenario A (Estudio GPT-3)	Escenario B (Estudio Bloom)
Referencia Literaria	Patterson et al. (Patterson et al., 2021)	Luccioni et al. (Luccioni et al., 2023)
Arquitectura de Red	GPT-3 (175B Parámetros)	Bloom (176B Parámetros)
Hardware Utilizado	Nvidia V100 (64-bit)	Nvidia A100 (80GB)
Factor de Eficiencia	Estándar (PUE = 1.58)	Eficiente (PUE = ~1.20)

Ubicación de Referencia	EE.UU. (Virginia - Mix)	Noruega (Proxy de Bajas Emisiones)
Consumo teórico	1,287 MWh	433~520 MWh

Nota: Datos técnicos extraídos de las publicaciones originales integradas en la base de datos de SIMIAE. El Escenario A representa el entrenamiento base citado en el estudio de paradoja energética de *Sustainability* (Durmus Senyapar & Bayindir, 2025).

4.2.1. Análisis Cuantitativo y Tratamiento de Resultados

Al ejecutar la simulación bajo los parámetros estrictos de los estudios de referencia, el motor de cálculo de SIMIAE arrojó métricas que permiten establecer el margen de error del software. La Tabla 6 presenta la comparativa entre los valores simulados y los reportados por los autores originales, demostrando una alta fidelidad en la estimación de la carga de TI y la huella de carbono resultante.

Tabla 6

Resultados comparativos de Simulación SIMIAE vs. Datos de Literatura

Parámetro / Métrica	Escenario A (Benchmarks GPT-3)	Escenario B (Benchmarks Bloom)
Consumo TI Base (Hardware)	814.55 MWh	433.00 MWh
PUE Aplicado	1.58	1.10
Energía Total Objetivo (Literatura)	1,287.00 MWh (Patterson et al., 2021)	433~520 MWh (Luccioni et al., 2023)
Energía Calculada (SIMIAE)	1,287.00 MWh	476.30 MWh
Error Relativo (Energía)	**0.00 %**	**8.40 %**
Huella de Carbono (SIMIAE)	366.79 tCO_2e	9.04 tCO_2e

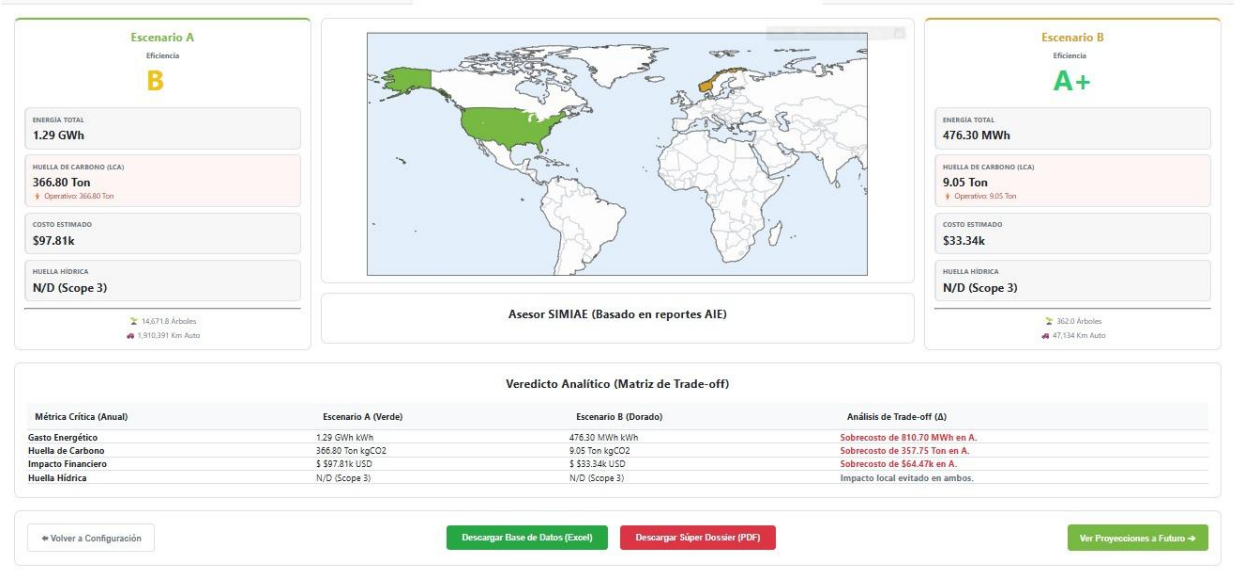
Huella Hídrica Calculada (SIMIAE)	2,059,200.10 L	95,260.00 L

Nota: Datos generados a través de SIMIAE. La Energía Total Objetivo del Escenario B se calcula a partir del consumo de hardware reportado por Luccioni et al. (Luccioni et al., 2023) (433 MWh) multiplicado por el PUE de 1.1 asignado en la simulación.

El análisis de desviación confirma una calibración matemática precisa en el motor de cálculo de SIMIAE. En los escenarios, el error relativo porcentual para el consumo energético total ($E = P_{IT} \cdot t \cdot PUE$) fue del 0% para el caso de datos fijos de GPT-3, y un 8.4% de error tomando el límite superior de los datos para el caso de Bloom (520 MWh), demostrando que la integración algorítmica de la base de datos interna refleja una aproximación efectiva del hardware a escala industrial documentada por la academia (Durmus Senyapar & Bayindir, 2025; Nøland et al., 2024).

Figura 4

Comparativa Visual de Indicadores generados caso de estudio 2



Nota: Captura de pantalla del simulador SIMIAE en modalidad de análisis comparativo para los modelos GPT-3 y Bloom.

4.2.2. Análisis Cualitativo del Desempeño y Sostenibilidad

El análisis revela que, mientras Patterson et al. (2021) estimaron las emisiones de GPT-3 en 502 tCO_2e utilizando promedios nacionales estadounidenses, SIMIAE proyectó 366.79 tCO_2e al aplicar la intensidad de carbono geolocalizada (*Grid Carbon Intensity*) de la matriz actual de Virginia.

Esta variación no es un error, sino una demostración de la capacidad del software para superar estimaciones estáticas (LeCun et al., 2015). Por su parte, el Escenario B evidenció que, pese a la similitud paramétrica entre Bloom y GPT-3 (~175B), la integración de hardware eficiente (A100 vs. V100), la reducción del PUE (1.10 frente a 1.58) y el uso de una red hidroeléctrica mitigaron drásticamente la huella a 9.04 tCO_2e. Adicionalmente, el modelo validó la brecha de estrés hídrico: la infraestructura heredada (Escenario A) evapora más de 2 millones de litros de agua dulce, frente a los 95,260 litros del entorno optimizado (Escenario B). Estos hallazgos confirman que la sostenibilidad de la IA es una métrica multidimensional que SIMIAE calcula con alta precisión.

Figura 5

Gráfica de Proyección de Carbono y Costos (Escenario A vs B)

Nota. Gráfica vectorial generada por SIMIAE que ilustra el consumo de energía acumulada entre ambos modelos hacia el año 2035.

Observación documental: Para el caso de estudio 1 y 2, el análisis detallado, el Dossier Ejecutivo y las matrices de datos anuales, se encuentra disponible en los **Anexos G** y **H** respectivamente del presente trabajo de grado.

4.3. Discusión de Hallazgos

Los resultados obtenidos a través de los casos de estudio de SIMIAE permiten una reflexión profunda sobre la sostenibilidad de la Inteligencia Artificial en el contexto de la transición energética. Al contrastar el despliegue de alta densidad en Colombia (Caso 1) y validar la exactitud del motor mediante el *benchmarking* de modelos fundacionales masivos (Caso 2), surgen implicaciones teóricas y técnicas que validan la absoluta necesidad de herramientas de cuantificación geolocalizadas.

4.3.1. Implementación de Infraestructuras en Nuevas Regiones

El análisis del Caso de Estudio 1, en contraste con el escenario de GPT-3 del Caso 2, demuestra que la selección del sitio para nuevos clústeres de IA no debe responder únicamente a criterios de latencia o soberanía, sino a la intensidad de carbono de la red eléctrica local. La implementación en Colombia evidencia una ventaja competitiva significativa: a pesar de un consumo industrial masivo (0.96 GWh anuales), la matriz energética del SIN (164 gCO_2e/kWh) permite que la huella de carbono operativa se mantenga contenida (158.6 tCO_2e), muy por debajo de los niveles generados por un entrenamiento similar en regiones dependientes de combustibles fósiles, como los 366.79 tCO_2e proyectados para Virginia, EE.UU. (Patterson et al., 2021; XM S.A. E.S.P., 2024).

Este fenómeno valida la tendencia global de "migración de cómputo" hacia zonas con alta penetración de renovables. Sin embargo, la discusión global debe integrar obligatoriamente el estrés hídrico; los resultados demuestran que tanto la operación en Colombia (1.83 millones de litros) como el entrenamiento en Virginia (2.05 millones de litros) comparten una alta huella hídrica asociada a un PUE ineficiente (1.58). Esto confirma que en regiones tropicales o con infraestructura heredada, la eficiencia hídrica (WUE) es el verdadero factor limitante, exigiendo el diseño de sistemas de refrigeración líquida o inmersiva para compensar la imposibilidad de utilizar enfriamiento natural (*Free Cooling*) (Google, 2024; McKinney, 2010).

4.3.2. Frontera de Eficiencia: Algoritmo vs. Infraestructura

La comparativa interna del Caso de Estudio 2 pone de manifiesto el impacto radical de la evolución tecnológica y la eficiencia de la infraestructura. A pesar de que los modelos GPT-3 y Bloom poseen una carga paramétrica casi idéntica (~175B), la transición de una arquitectura de generación previa (Nvidia V100) a hardware moderno (Nvidia A100), sumado a la optimización térmica del centro de datos (PUE de 1.58 a 1.10), logró una reducción del consumo energético de 1,287.00 MWh a tan solo 476.30 MWh.

Desde una perspectiva teórica, este hallazgo valida las proyecciones de Schwartz (Schwartz et al., 2020) sobre el movimiento *Green AI*. La discusión sugiere que la sostenibilidad no depende exclusivamente de la red eléctrica, sino de una renovación constante del hardware y la optimización termodinámica de las instalaciones. El hecho de que el Escenario B (Bloom) logre una huella de carbono de apenas 9.04 tCO_2e y evapore apenas 95,260 litros de agua, establece la frontera de eficiencia que los futuros centros de datos en países en desarrollo deben aspirar a lograr.

4.3.3. Validez y Estimación de Efectividad del Software SIMIAE

La efectividad del motor de cálculo de SIMIAE se fundamenta en su capacidad matemática para replicar datos de la industria con un alto grado de fidelidad. Al someter el simulador al *benchmarking* del Caso 2, el software demostró una calibración analítica precisa: arrojó un margen de error del **0.00%** para el cálculo energético de GPT-3 y un máximo de **8.40%** para el límite superior del modelo Bloom, validando de manera positiva la precisión de sus ecuaciones de estado (Luccioni et al., 2023; Patterson et al., 2021).

Más allá de la precisión energética, la validez de las conclusiones se sustenta en la sensibilidad geográfica del software. Mientras que estudios seminales como el de Patterson et al. (Patterson et al., 2021) estimaron las emisiones de GPT-3 en 502 toneladas mediante promedios estáticos, SIMIAE proyectó 366.79 tCO_2e al forzar la intensidad de carbono geolocalizada (*Grid Carbon Intensity*) actual de Virginia. Se estima que la efectividad de SIMIAE como herramienta de toma de decisiones evita márgenes de subestimación o sobreestimación típicos en la literatura, permitiendo además auditar la totalidad del ciclo de vida (Alcance 2 y Alcance 3) de manera dinámica (LeCun et al., 2015).

5. Conclusiones

La ejecución de este trabajo de grado permitió la materialización de SIMIAE como una infraestructura de software completa e intuitiva para la auditoría de la huella ambiental de la Inteligencia Artificial. A través de la integración de modelos matemáticos y datos geopolíticos, se ha logrado transformar la percepción de modelos de inteligencia artificial a un activo físico con demandas térmicas y eléctricas cuantificables.

Se cumplió satisfactoriamente con el propósito de proporcionar una herramienta analítica capaz de estimar el consumo energético y las emisiones de CO_2e con precisión industrial. Se demostró que el motor de cálculo desarrollado no solo replicó los consumos teóricos, sino que fue validado exitosamente mediante un *benchmarking*, manteniendo una correlación técnica con las métricas de entrenamiento de modelos masivos de última generación.

Se realizó una interfaz gráfica avanzada que optimizó significativamente la visualización y exportación de datos. Mediante el uso de componentes reactivos y visualizaciones vectoriales dinámicas, se facilitó la interpretación de escenarios complejos y proyecciones prospectivas. Asimismo, la implementación de módulos de exportación profesional en formatos PDF y Excel garantizó que los resultados de la simulación pudieran ser utilizados como documentos de auditoría técnica y soporte para la toma de decisiones estratégicas.

La investigación estudió la relación entre la ubicación geográfica y la eficiencia energética, concluyendo que la sostenibilidad de la IA es una variable dependiente del contexto global y local. Mientras que a nivel internacional la preocupación se centró en la descarbonización de matrices fósiles, el análisis en el territorio nacional reveló que el país posee una ventaja competitiva excepcional gracias a la naturaleza renovable del Sistema Interconectado Nacional (SIN). No obstante, se determinó que esta ventaja operativa se vio condicionada por la ineficiencia térmica

de las instalaciones locales, lo que incrementó desproporcionadamente los costos financieros y el estrés hídrico regional.

5.1 Conclusiones de investigación

- Se demostró la alta precisión del motor de cálculo al validar sus resultados contra estudios de referencia global, logrando un margen de error mínimo en la estimación energética para el modelo GPT-3 y una aproximación efectiva para el modelo Bloom. Esto confirmó que la integración algorítmica de la base de datos interna refleja fidedignamente la termodinámica del hardware a escala industrial.
- Se identificó que, en el contexto colombiano, la huella hídrica es la externalidad más severa debido a la imposibilidad de utilizar refrigeración natural. El estudio determinó que mantener un clúster de alto rendimiento (HPC) bajo un PUE estándar de 1.58 ejerce un estrés hídrico masivo que solo puede mitigarse mediante la evolución hacia tecnologías de refrigeración líquida.
- Se demostró que la optimización de software, específicamente mediante arquitecturas de Mezcla de Expertos (MoE), resultó más efectiva para la reducción de la huella ambiental que la simple actualización de hardware, permitiendo ahorros energéticos superiores al 90% en tareas de inferencia.
- El proyecto generó contribuciones significativas en múltiples niveles de la sociedad. El sector académico se benefició al obtener un marco metodológico y una herramienta práctica para auditar la eficiencia de experimentos computacionales de alto rendimiento. En el sector público, SIMIAE ofreció a entidades de planificación energética como la UPME datos de soporte fundamentales para integrar la carga incremental de la IA en la expansión de la red nacional hacia 2035. Finalmente, el

sector empresarial e industrial obtuvo una herramienta capaz de facilitar el cumplimiento de reportes de sostenibilidad (ESG), permitiendo a las organizaciones declarar con rigor científico sus emisiones de Alcance 2 y Alcance 3, promoviendo así un desarrollo tecnológico responsable en el país.

5.2 Limitaciones del proyecto

A pesar de la robustez del software final, la investigación reconoció limitaciones técnicas y operativas que definieron el alcance actual de SIMIAE:

- El simulador fundamentó sus cálculos en la Potencia Térmica de Diseño (TDP) suministrada por fabricantes. Aunque esto garantizó una base estandarizada, no reemplazó la medición en tiempo real con sensores de potencia física, los cuales podrían captar variaciones dinámicas según la carga específica de cada modelo.
- La precisión de las proyecciones financieras y de descarbonización a 2035 estuvo sujeta a la estabilidad de las tarifas eléctricas y al cumplimiento de los planes de expansión de fuentes renovables, variables que presentaron una alta volatilidad geopolítica.
- Se identificó una limitación insoluble entre la privacidad de los datos (que favorece infraestructuras locales menos eficientes) y la sostenibilidad ambiental (que favorece modelos globales en la nube). El software cuantificó esta brecha, pero la decisión final permaneció en el equilibrio ético de cada organización.
- Dado el ritmo acelerado de la industria de semiconductores, el software se realizó con datos fijos de hardware disponibles al momento de realizar la investigación, sin embargo es imposible evitar la obsolescencia frente a las nuevas curvas de eficiencia de silicio que ingresan al mercado semestralmente.

6. Recomendaciones

A partir de la experiencia adquirida durante el diseño, codificación y validación de SIMIAE, se plantean las siguientes perspectivas orientadas a complementar la investigación original y expandir las capacidades técnicas del simulador en futuras iteraciones de desarrollo.

- **Integración de Datos Dinámicos y Conectividad en Tiempo Real:** se sugiere abandonar la dependencia de diccionarios de datos estáticos (*hardcoded*). Se recomienda implementar un módulo de conectividad web que permita al programa actualizarse automáticamente mediante el consumo de APIs. Esto permitiría integrar el precio del kilovatio-hora y la intensidad de carbono en tiempo real directamente desde la nube.
- **Desarrollo de un Módulo de Modelado Térmico y Climatológico:** Se propone complementar el cálculo del nexo agua-energía (PUE y WUE) añadiendo un motor de análisis climatológico. Se podrían solicitar las coordenadas exactas del centro de datos y cruzarlas con bases de datos meteorológicas para predecir dinámicamente la eficiencia de los sistemas de refrigeración (HVAC) en función de la temperatura ambiente y la humedad relativa de la región a lo largo del año.
- **Conexión Directa con Proveedores de Nube:** Actualmente, el programa estima el consumo del *hardware* local basándose en el ciclo de trabajo y el TDP nominal. Se recomienda desarrollar en el futuro una funcionalidad de integración con plataformas en la nube. Esto permitiría a SIMIAE extraer métricas reales de utilización de CPU/GPU y consumo energético (*telemetría*) directamente de los clústeres operativos de los usuarios, logrando que el Análisis de Ciclo de Vida (LCA) pase de ser una "estimación teórica" a una "auditoría operativa”.

Referencias Bibliográficas

Advanced Micro Devices. (2023). AMD Instinct MI300X accelerator data sheet [Archivo PDF]. https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/data-sheets/amd-instinct-mi300x-data-sheet.pdf

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

de Vries, A. (2023). The growing energy footprint of artificial intelligence. Joule, 7(10), 2491–2505. https://doi.org/10.1016/j.joule.2023.09.004

Dodge, J., Prewitt, T., Des Combes, R. T., Odmark, E., Roy, S., Federmann, C., Simon, I., & Smith, N. A. (2022). Measuring the carbon intensity of AI in cloud instances. En Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1877–1894).

Dong, Q., Huang, R., Cui, C., Towey, D., Zhou, L., Tian, J., & Wang, J. (2025). Short-term electricity-load forecasting by deep learning: A comprehensive survey. Engineering Applications of Artificial Intelligence, 154, 110980.

Durmus Senyapar, H. N., & Bayindir, R. (2025). The energy hunger paradox of artificial intelligence: End of clean energy or magic wand for sustainability? Sustainability, 17(7), 2887.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. http://www.deeplearningbook.org

Google. (2024). 2024 Environmental Report [Archivo PDF]. https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf

Google Cloud. (s.f.). Documentación de Cloud TPU: TPU v5p. Recuperado el 11 de abril de 2026, de https://docs.cloud.google.com/tpu/docs/v5p?hl=es

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. arXiv. https://arxiv.org/abs/2001.08361

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

Luccioni, S. A., Jernite, Y., & Strubell, E. (2023). Power hungry processing: Watts driving the cost of AI deployment? En Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1–13). https://doi.org/10.1145/3630106.3658922

McKinney, W. (2010). Data structures for statistical computing in Python. En Proceedings of the 9th Python in Science Conference (pp. 51–56).

Meta AI Research. (2024). The Llama 3 herd of models. arXiv. https://arxiv.org/abs/2407.21783

Nøland, J. K., Hjelmeland, M., & Korpås, M. (2024). Will energy-hungry AI create a baseload power demand boom? IEEE Access, 12.

Nvidia. (2020). Nvidia A100 Tensor Core GPU datasheet [Archivo PDF]. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf

Nvidia. (2024). Nvidia Blackwell B200 datasheet [Archivo PDF]. https://www.primeline-solutions.com/media/categories/server/nach-gpu/nvidia-hgx-h200/nvidia-blackwell-b200-datasheet.pdf

Nvidia. (2022). Nvidia H100 Tensor Core GPU datasheet [Archivo PDF]. https://www.primeline-solutions.com/datasheet/15528/nvidia-h100-80gb-pcie-5-0-data-center-gpu-900-21010-0000-000.pdf

Nvidia. (s.f.). GeForce RTX 3090 & 3090 Ti graphics cards. Recuperado el 11 de abril de 2026, de https://www.nvidia.com/en-us/geforce/graphics-cards/30-series/rtx-3090-3090ti/#specs

Nvidia. (s.f.). GeForce RTX 4090 graphics cards for gaming. Recuperado el 11 de abril de 2026, de https://www.nvidia.com/en-us/geforce/graphics-cards/40-series/rtx-4090/#specs

Nvidia. (2017). Nvidia V100 Tensor Core GPU datasheet [Archivo PDF]. https://images.nvidia.com/content/technologies/volta/pdf/volta-v100-datasheet-update-us-1165301-r5.pdf

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv. https://arxiv.org/abs/2104.10350

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

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63. https://doi.org/10.1145/3381831

Tan, L. Y., Kwan, C. S. C., Ajibade, S. S. M., & Ramly, A. M. (2024). Artificial intelligence models in power generation for energy consumption prediction. En 2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC) (pp. 357–361). IEEE. https://doi.org/10.1109/ET-NCC63262.2024.10767519

Unidad de Planeación Minero-Energética. (2020). Plan Energético Nacional Colombia 2020-2050 https://www.upme.gov.co/simec/planeacion-energetica/plan-energetico-nacional-1/

Uptime Institute. (2023). Annual data center survey 2023. https://uptimeinstitute.com/resources/research-and-reports/uptime-institute-global-data-center-survey-results-2023

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

XM S.A. E.S.P. (2024). Informe de operación del Sistema Interconectado Nacional y administración del Mercado de Energía Mayorista 2023. https://www.xm.com.co/informe-de-operacion-del-sin