Ciencia Latina Revista Científica Multidisciplinar, Ciudad de México, México.
ISSN 2707-2207 / ISSN 2707-2215 (en línea), noviembre-diciembre 2026,
Volumen 10, Número 6.

https://doi.org/10.37811/cl_rcm.v10i6

ANÁLISIS DE EFICIENCIA ENERGÉTICA PARA MODELOS DE APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING) Y APRENDIZAJE PROFUNDO (DEEP LEARNING): UN ENFOQUE COMPARATIVO MULTIMODELO ESTANDARIZADO.

ENERGY EFFICIENCY ANALYSIS FOR MACHINE LEARNING AND DEEP LEARNING MODELS: A STANDARDIZED MULTI-MODEL COMPARATIVE APPROACH.

Esteban Fernando Castillo Durán
Universidad Americana de Europa UNADE

Rodrigo Cadena Martínez
Universidad Americana de Europa UNADE

DOI: https://doi.org/10.37811/cl_rcm.v10i6.26852

Análisis de eficiencia energética para modelos de Aprendizaje Automático (*Machine Learning*) y Aprendizaje Profundo (*Deep Learning*): un enfoque comparativo multimodelo estandarizado.

Esteban Fernando Castillo Durán.[1]
esfecasdu@gmail.com
https://orcid.org/0009-0007-4302-1708
Universidad Americana de Europa UNADE México

Rodrigo Cadena Martínez.
rodrigo.cadena@unade.edu.mx
https://orcid.org/0000-0001-9323-6132
Universidad Americana de Europa UNADE México

RESUMEN

Debido a la masificación en el consumo de tecnologías de inteligencia artificial, la productividad humana, tanto en el campo individual, empresarial, corporativo y colaborativo se ha beneficiado de un aumento notable en su aplicabilidad, a medida que estas herramientas han evolucionado en los últimos años, es importante tomar en consideración los factores que su utilización conlleva en múltiples ámbitos, entre los cuales se encuentra la sostenibilidad y el impacto ambiental, es innegable el hecho que el progreso de estos sistemas está asociado a un notable incremento en el consumo de energía eléctrica, lo cual acarrea grandes desafíos para su correcta gestión. El presente documento comprende el estudio de eficiencia energética aplicada al campo de la Inteligencia Artificial, se analiza varios modelos o paradigmas comúnmente utilizados por empresas en sus actividades diarias, sometiéndoles a pruebas realizadas en entornos controlados y bajo condiciones idénticas, con la finalidad de evaluar su rendimiento en materia energética, lo cual ha permitido realizar una comparativa entre estos, y de esta manera poder indicar objetivamente cuáles son los más eficientes en términos de consumo energético.

Palabras clave: Inteligencia Artificial, consumo energético de la I.A., impacto ambiental en modelos de Aprendizaje Automático – Aprendizaje Profundo.

[1] Autor principal
Correspondencia: esfecasdu@gmail.com

Energy Efficiency Analysis for Machine Learning and Deep Learning Models: A Standardized Multi-Model Comparative Approach.

ABSTRACT

Due to the widespread adoption of artificial intelligence technologies, human productivity—across individual, business, corporate, and collaborative contexts—has seen a notable boost. As these tools have evolved in recent years, it is essential to consider the implications of their use across various domains, including sustainability and environmental impact. It is undeniable that the advancement of these systems is linked to a significant increase in electricity consumption, which presents major challenges for effective management. This document examines energy efficiency within the artificial intelligence field. It analyzes several models or paradigms commonly used by companies in their daily operations, subjecting them to tests in controlled environments under identical conditions to evaluate their energy performance. This allows for a comparison between the models, making it possible to objectively identify those that are the most efficient in terms of energy consumption.

Keywords: Artificial Intelligence, AI energy consumption, environmental impact of Machine Learning – Deep Learning models.

Artículo recibido 10 noviembre 2026
Aceptado para publicación: 17 noviembre 2026

INTRODUCCIÓN.

El avance en términos tecnológicos que tienen los sistemas de Inteligencia Artificial (*I.A.*), se ha producido a un ritmo tan exponencial y acelerado, que muy difícilmente es comparable con cualquier otro tipo de tecnologías que hayan presentado avances notables en la última década, tal es el caso que, si contrastamos los avances producidos postpandemia en este campo, han ocurrido en períodos de tiempo tan cortos que, incluso el seguir el ritmo de las nuevas aplicaciones se ha convertido en un reto para muchos sectores. No obstante, todas estas nuevas herramientas sin excepción hacen uso de determinados modelos de I.A. los cuales, hablando ya en términos técnicos conforman la columna vertebral, el (*backend*) de todas las herramientas y plataformas más populares a día de hoy.

Justificación técnica y contexto del problema.

Existe una problemática muy grande que se deriva del uso de las tecnologías de IA, el impacto ambiental producido por los centros de datos de IA, tal y como lo menciona Tangermann, V. (2026) sobre la investigación realizada por científicos de Cornell, quienes estimaron que la tasa de crecimiento en infraestructura de I.A. representaría en términos de huella ambiental un total de 24 a 44 millones de toneladas métricas de dióxido de carbono para el año 2030, cifra equiparable a incorporar un promedio entre cinco a diez millones de vehículos al parque automotor de Estados Unidos.

La infraestructura digital también ha necesitado actualizarse al mismo ritmo para permitir albergar en su hardware a dichas tecnologías, por lo tanto, la huella generada de recursos tanto computacionales, energéticos e hídricos inclusive, han elevado el consumo de recursos hasta límites históricos sin precedentes de acuerdo con Aczel et al. (2026). Este escenario ha motivado la revisión crítica del desarrollo tecnológico en este campo del conocimiento.

Tabla 1: Delimitación de la investigación.

Pregunta de investigación:	¿De qué manera varía el consumo energético entre los distintos paradigmas algorítmicos de *Machine Learning* y *Deep Learning*?
Objetivos:	General. Evaluar y comparar de una manera empírica el consumo de energía eléctrica, la potencia sostenida, las emisiones de carbono y la eficiencia unitaria normalizada para un conjunto de familias de modelos y arquitecturas de aprendizaje automático (*M.L.*) y aprendizaje profundo (*D.L.*). Específicos. ▪ Analizar el comportamiento energético tanto en potencia media sostenida (*W*), como el consumo acumulado (*kWh*), en aceleradores de hardware tipo GPU, también en unidades CPU bajo condiciones de ejecución continua. ▪ Formular e integrar métricas unitarias normalizadas por trabajo útil (*µWh/muestra*, *mWh/época*, *Wh/1k tokens*, *Wh/hora de audio*, *µWh/consulta*), que permitan realizar una comparación objetiva dentro de las distintas arquitecturas. ▪ Evaluar el impacto energético para tres paradigmas diferenciados que fueron evaluados: la carga repetitiva en CPU, el entrenamiento real por épocas en GPU y la inferencia continua/sostenida en arquitecturas modernas. ▪ Identificar y documentar brechas o subestimaciones en la medición telemétrica.
Alcance del Estudio:	La ejecución de este estudio comprende los siguientes dominios: ▪ Espectro tecnológico: mediante la evaluación de 16 paradigmas algorítmicos clasificados cronológicamente y divididos en cuatro subsistemas: *(a) Aprendizaje Tabular Clásico, (b) Deep Learning y NLP, (c) Visión Avanzada y Multimodalidad, (d) Modelos Generativos, Inferencia Agentica y arquitecturas RAG.* ▪ Protocolo de Medición: ejecución de 51 experimentos de 60 minutos continuos sumando un total de 51 horas de cómputo neto supervisado, se excluye la medición PUE (*Power Usage Effectiveness*), ya que, al estar asociado explícitamente al centro de datos es un valor fuera del alcance de control debido a la asignación aleatoria de entornos de ejecución. ▪ Infraestructura empleada: Servidores de Google Colab equipados con hardware específico, monitoreados continuamente en intervalos de 10 minutos por medio de la librería CodeCarbon.

Analizando el contexto actual, se puede determinar que el despliegue de Modelos de Lenguaje de Gran Escala (*LLMs*), ha ubicado directamente a la infraestructura de centros de datos especializados en el centro del debate socioambiental contemporáneo, por un lado la postura corporativa resalta el potencial transformador de estas tecnologías justificándolas con la optimización de procesos y la aceleración de resultados y por tanto de conocimiento, pero por otro, el sustento material sobre el que se desarrollan estas actividades, exige una cantidad de recursos que cada vez es más elevada, lo que se traduce en una demanda en constante crecimiento.

Bajo este contexto entra en escena la aparición de dos conceptos fundamentales, IA Roja (*Red AI*), y la IA Verde o Ecológica (*Green AI*), con la finalidad de categorizar las prioridades metodológicas de la investigación sobre aprendizaje profundo (Schwartz et al., 2020).

Dicha investigación describe a la I.A. Roja como una tendencia en la cual el éxito se evalúa de manera casi exclusiva mediante métricas de precisión (*accuracy*), puntuaciones de referencia, o estándares (*benchmarks*); por tanto para este enfoque el consumo energético, la infraestructura de procesamiento y el volumen de datos son considerados como variables ilimitadas o externalidades sin costo relevante, operando bajo la lógica de escalado exponencial que exhibe marcados rendimientos decrecientes, para lograr mejoras marginales en la precisión, el volumen de cómputo requerido se multiplica por factores de 10x a 100x, traduciéndose en un crecimiento del procesamiento empleado en los entrenamientos avanzados de más de 300,000 veces entre 2012 y 2018 (Schwartz et al., 2020).

En términos energéticos el entrenamiento de un modelo Transformador (*Transformer*) puede generar aproximadamente 284 toneladas de dióxido de carbono equivalente (CO_2e), cifra que supera cinco veces las emisiones del ciclo de vida de un automotor promedio, según el estudio de (Strubell et al., 2019).

La evaluación de modelos frontera acentúa esta magnitud, la Universidad de las Naciones Unidas, en su investigación detalla que, el entrenamiento de GPT-4 requirió entre 50 y 70 *GWh* de electricidad durante 100 días, generando una huella de carbono de 25,000 toneladas de CO_2e y un consumo directo de 592 a 600 millones de litros de agua, equivalente a la capacidad de 237 piscinas olímpicas (UNU-INWEH, 2026). Por ende, las proyecciones para modelos subsecuentes como GPT-6 contemplan requerimientos eléctricos de entrenamiento de hasta 100 *GWh* y emisiones de 42,000 toneladas de CO_2e, reafirmando la insostenibilidad del modelo de IA Roja si no se imponen restricciones de eficiencia.

Frente a esta panorámica insostenible, la I.A. Verde (*Green AI*) propone reconceptualizar la eficiencia computacional y ambiental como la métrica de primer orden, indicando que debería ir a la par con la precisión algorítmica (Schwartz et al., 2020). Una de las características más notables es promover la reducción de la huella ecológica a través de técnicas de optimización, del uso de arquitecturas eficientes y la transparencia en el reporte de datos, garantizando que los beneficios sean viables a largo plazo sin comprometer la estabilidad ecológica ni la equidad distributiva de recursos naturales.

Además del análisis energético, este modelo aplica evaluaciones del ciclo de vida (*LCA*) que añaden información sobre la infraestructura, las investigaciones de Luccioni et al., (2023); Wu et al., (2022) demostraron mediante el análisis del modelo de lenguaje BLOOM (con 170 millones de parámetros) que de su huella total de 50.5 toneladas de CO_2e, el 48.9 % correspondió al consumo dinámico de las GPUs, el 28.9 % al consumo inactivo y la infraestructura del centro de datos, y un 22.2 % a la fabricación y transporte del hardware (Luccioni et al., 2024).

Según Mistral AI (2025) la cuantificación rigurosa del impacto ambiental requiere superar las estimaciones teóricas simplificadas y adoptar marcos metrológicos estandarizados que abarquen la totalidad de la cadena operativa, tradicionalmente la industria se centró en la "Eficiencia en el uso de la Energía", conocido como el factor PUE, con un valor cercano a 1.0, aunque no deseable, no refleja el consumo absoluto, para resolver estas limitaciones la investigación realizada por Lin et al. (2023), establece las directrices de la industria que proponen marcos de métricas estructurados en cinco categorías: energía, emisiones de gases efecto invernadero (*GEI*), agua, residuos y ecosistema local evaluados en una pirámide de tres niveles de madurez organizacional.

Esta visión holística permite monitorear desde los indicadores básicos de consumo eléctrico hasta las evaluaciones de desempeño hídrico (*Water Usage Effectiveness - WUE*) e intensidad de carbono en tiempo real, garantizando un reporte transparente y comparable frente a regulaciones internacionales.

Mediante la instrumentación directa de la infraestructura de producción para el asistente Gemini, la metodología *full-stack* demostró que una consulta de texto mediana consume 0.24 *watt-horas* (*Wh*) de energía, genera 0.03 gramos de CO_2e y utiliza 0.26 mililitros de agua, en términos comparativos cotidianos el consumo energético por consulta, equivale a mirar televisión durante aproximadamente nueve segundos, ubicándose sustancialmente por debajo de las estimaciones públicas que sugerían consumos entre 10 y 50 mL. de agua por interacción (Elsworth et al., 2025).

El reporte ambiental de Google 2026, categoriza las emisiones en tres ejes operativos, el "*Alcance 1*" abarca las emisiones directas generadas por fuentes bajo control de la entidad (combustión de diésel en generadores y pérdidas de refrigerantes), el "*Alcance 2*", cuantifica las emisiones indirectas asociadas a la generación de electricidad, vapor o refrigeración adquiridos, por último el "*Alcance 3*" engloba todas

las demás emisiones indirectas a lo largo de la cadena de valor (incluyendo extracción de materias primas, fabricación de servidores y construcción de centros de datos). El mismo reporte señala que, en 2025 el consumo total de toda la infraestructura de Google, alcanzó los 44,046,500 *MWh* (44 *TWh*), registrando un crecimiento del 37% en su carga totalizada; a escala macroeconómica, el consumo global de centros de datos alcanzó los 448 *TWh* en 2025, en Estados Unidos, regiones como el estado de Texas, se ha convertido en el epicentro de expansión de la I.A., diversas empresas explotan vacíos legales en las licencias ambientales para instalar infraestructura necesaria aprovechándose de esta situación.

Por ende, el impacto ambiental de los centros de datos se extiende de manera crítica hacia los recursos hídricos y el microclima de las comunidades anfitrionas; a nivel global un estudio cuantitativo sobre 9,347 instalaciones reveló que el 25% de los centros de datos se ubican en zonas que ya experimentan al menos un mes de déficit hídrico (*water gap*) al año bajo condiciones climáticas de referencia (Rosa et al., 2026). Los mismos autores afirman que, se ha presentado un conflicto de superposición espacial con otros usos prioritarios del suelo y agua, el 41% de los centros de datos globales (3,582) se localizan en áreas dominadas por la demanda municipal de agua potable, mientras que el 28% (2,573) colindan con zonas de agricultura y regadío; un 18% en el sector manufacturero, así por ejemplo en Estados Unidos, más del 54% de los centros de datos compiten directamente con redes de suministro urbano y un 20% con actividades agrícolas, acentuando el riesgo de desabastecimiento en estados del oeste y sur del país en períodos de sequía.

Además del estrés hídrico, el estudio dirigido por Marioni et al. (2026) concluyó que la disipación masiva del calor generado por estos centros, producen una alteración de microclima local, efecto conocido como "Isla de Calor (*Heat Island Effect*)", un ejemplo de esta situación se detectó en el complejo industrial de Slough, que comprende la mayor concentración de centros de datos de Europa con casi un gigavatio de potencia instalada, gracias a investigaciones satelitales de la Universidad de Cambridge, se confirmó que la densidad de las torres de enfriamiento y sistemas de ventilación eleva la temperatura ambiente de la zona circundante entre 2 °*C* y 9 °*C*, generando condiciones sofocantes para la población residente.

Esta situación ha desencadenado una ola de resistencia social sin precedentes, transformando disputas vecinales aisladas en un movimiento político-activista coordinado, según reportes de la organización

Data Center Watch, el número de grupos comunitarios en oposición organizada se duplicó en Estados Unidos en el primer trimestre de 2026, pasando de 396 a 833 grupos distribuidos en varios estados. Para reconciliar la expansión tecnológica con los límites ecológicos y sociales, la comunidad científica avanza en propuestas que permitan mitigar con soluciones integrales, entre las que están la transición de refrigeración por aire hacia sistemas de refrigeración líquida directa al chip, que permite disipar eficientemente densidades térmicas superiores a los 100 *kW* por rack en GPUs y aceleradores avanzados (Carneiro et al., 2018).

Otra solución técnica de vanguardia es la refrigeración líquida mediante fluidos dieléctricos con capacidad calorífica sustancialmente superior al aire, logrando reducir el consumo eléctrico dedicado al enfriamiento hasta en un 10% y permitiendo alojar hasta el doble de capacidad informática en la misma superficie física. Paralelamente, las estrategias de suministro eléctrico de avanzada se orientan hacia la adquisición de energía libre de Carbono las 24 horas del día (*24/7 CFE*), mediante acuerdos de compra de energía (*Power Purchase Agreements - PPAs*), y tecnologías de generación limpia, adicionalmente y según detallan en su investigación Xiao et al., (2026) se está acuñando el concepto de computación consciente del carbono (*Carbon-Aware Computing*), que se basa en reubicar dinámicamente las cargas de trabajo diferibles hacia centros de datos ubicados en regiones geográficas con abundancia temporal de energía renovable.

No obstante, estas soluciones enfrentan la Paradoja de Jevons (*o efecto rebote*), este principio establece que, los incrementos en la suficiencia de un recurso tienden a aumentar, en lugar de reducir su consumo total, al abaratar los costos de producción, esto se podría ejemplificar bajo la premisa que la reducción dramática en el consumo de energía por token lograda por modelos optimizados en las arquitecturas más eficientes, reducen el costo por consulta, lo que detona en un crecimiento exponencial del volumen de usuarios y aplicaciones integradas a un determinado modelo.

METODOLOGÍA.

Se planteó un diseño de investigación con enfoque cuantitativo, cuasiexperimental y correlacional; fue estructurado sobre un protocolo de evaluación estandarizado que comprendió la ejecución y análisis de 16 paradigmas tecnológicos de inteligencia artificial, relacionados a los subcampos de Aprendizaje Automático (*M.L.*) y Aprendizaje Profundo (*D.L.*); el proceso evalúa la carga energética producida durante la ejecución continua de 60 minutos, realizando registros de muestreo cada 10 minutos por cada modelo estudiado, para garantizar la validez estadística y la repetibilidad de datos, cada uno de los paradigmas seleccionados se ejecutó en un triplicado experimental independiente (*ext RUN1, Ext RUN2, Ext RUN3*).

Tabla 2: Matriz cronológica de los paradigmas estudiados

Familia Tecnológica	Paradigma / Modelo	Año de Publicación	Autores Principales & Referencia Canónica (Fuente Primaria)
1. Aprendizaje Tabular y Clásico	Regresión Lineal	1805 / 1809	Adrien-Marie Legendre (1805) / Carl Friedrich Gauss (1809)
	KNN (*K-Nearest Neighbors*)	1951 / 1967	Evelyn Fix & Joseph Hodges (1951) / Thomas Cover & Peter Hart (1967)
	K-Means Clustering	1956 / 1967	Hugo Steinhaus (1956) / Stuart Lloyd (1957) / James MacQueen (1967)
	Random Forest Classifier	2001	Leo Breiman (2001)
	XGBoost Classifier	2016	Tianqi Chen & Carlos Guestrin (2016)
2. Deep Learning & NLP	Perceptrón Multicapa (MLP)	1958 / 1986	Frank Rosenblatt (1958) / David Rumelhart, Geoffrey Hinton & Ronald Williams (1986)
	Autoencoder (Neuronal)	1986 / 1988	David Rumelhart et al. (1986) / Hervé Bourlard & Yves Kamp (1988)
	LSTM (*Long Short-Term Memory*)	1997	Sepp Hochreiter & Jürgen Schmidhuber (1997)
	DistilBERT *Fine-Tuning*	2019	Victor Sanh, Lysandre Debut, Julien Chaumond & Thomas Wolf (2019)
3. Visión & Multimodalidad	ResNet-18 (CNN Deep)	2015	Kaiming He, Xiangyu Zhang, Shaoqing Ren & Jian Sun (2015/2016)
	YOLOv8n Object Detection	2016 / 2023	Joseph Redmon et al. (2016 - YOLOv1) / Glenn Jocher et al. (2023 - Ultralytics)
	Late Fusion Multimodal	2005 / 2020	Alex Graves & Jürgen Schmidhuber (2005) / Douwe Kiela et al. (2020)
4. Generativo, Audio, Agentes & RAG	DDPM Difusión Visual	2015 / 2020	Jascha Sohl-Dickstein et al. (2015) / Jonathan Ho, Ajay Jain & Pieter Abbeel (2020)
	Sistema RAG (SQuAD + FAISS)	2020	Patrick Lewis, Ethan Perez, Aleksandra Piktus et al. (2020)
	Faster-Whisper ASR	2022 / 2023	Alec Radford et al. / OpenAI (2022) / Guillaume Klein et al. (2023)
	Inferencia Agéntica (ReAct)	2022	Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du et al. (2022/2023)

***Nota**:* los 16 paradigmas del estudio se ordenaron según el año de publicación de cada algoritmo/modelo/paradigma, y se clasificaron en familias según su tipo.

Se determinó también la fijación estricta de semillas deterministas y pseudoaleatorias (*seed = 42*), esto responde a dos fundamentos metodológicos clave:

El *primero* tiene que ver con el "Propósito de Uso (*Garantía de Determinismo y Repetibilidad*)" ya que en algoritmos de estas ramas de la I.A., los procesos como el muestreo, la partición de datos, la inicialización de pesos, la ordenación aleatoria de lotes y el muestreo en optimizadores o decodificadores introduce variabilidad estocástica, por tanto al fijar el estado del generador de números pseudoaleatorios de manera idéntica para todas las librerías del ecosistema (*random, numpy, torch, torch.cuda*), se asegura que todos los experimentos ejecuten exactamente la misma secuencia de operaciones matemáticas; asegurando que las fluctuaciones observadas en el consumo energético (*kWh*) o la potencia (*W*) se atribuyan exclusivamente al comportamiento físico del hardware.

El *segundo* es la definición del valor desde un punto de vista matemático, para los generadores pseudoaleatorios (*PRNG*), cualquier número entero posee las mismas propiedades probabilísticas de dispersión, sin embargo, la elección del número 42 se adopta como estándar de transparencia técnica para alinearse con los valores por defecto utilizados por *Scikit-Learn*, facilitando la auditabilidad directa por parte de la comunidad académica.

Tabla 3: Sistema de variables y factores de confusión

Variable independiente	Paradigma analizado (16 en total)
Variables dependientes	Potencia Media Sostenida en Watts (*W*). Consumo energético acumulado en Kilovatios-hora (*kWh*). Huella de Carbono en gramos de dióxido de carbono equivalente (*g CO_2eq*). Métricas de calidad técnica de dominio (*Accuracy, F1-Score, ROC-AUC, mAP@0.5, RTF, Exact Match)* Eficiencia Normalizada por Unidad de Trabajo Útil
Variables controladas	Duración exacta del ciclo de ejecución (60 minutos). Intervalos de muestreo de Codecarbon (ciclos de 10 minutos). Entorno de ejecución en nube Google Colab Enterprise. Semilla determinista (seed=42).
Factores de confusión detectados	Heterogeneidad inherente a *datasets*. Diversidad en número de parámetros. Sobrecargas de los *frameworks*

Nota: para mitigar los factores de confusión, el estudio evitó categorizar entre dominios dispares valores de precisión, enfocándose en la intensidad eléctrica y eficiencia unitaria en cada caso.

Con respecto a la infraestructura, el entorno de cómputo se dividió según el tipo de procesamiento requerido por el algoritmo; por un lado, se contó con el entorno "CPU Base", destinado a los algoritmos tabulados de aprendizaje estadístico tradicional (Regresión Lineal, KNN, Random Forest, XGBoost y K-Means), ejecutados sobre procesadores Intel Xeon (2 vCPUs asignadas, 12,7 GB de RAM); también se presentó el "Entorno Acelerado GPU T4", destinado a paradigmas más recientes (arquitecturas de redes neuronales profundas, visión por computador, procesamiento de lenguaje natural, modelos generativos, audio ASR, inferencia agéntica y RAG), ejecutados sobre una GPU acelerador NVIDIA Tesla T4 (arquitectura Turing, 2,560 Tensor Cores, 16 GB VRAM GDDR6, TDP *70 W nominal / ~78 W sostenido*). Existe también un caso denominado "*Puente*" que se ejecutó sobre el paradigma *Autoencoder*, sirvió para cuantificar la diferencia de consumo y la aceleración relativa entre las plataformas de CPU y GPU respectivamente, sin alterar la arquitectura algorítmica ni el conjunto de datos de entrenamiento, ya que fue considerado el punto de inflexión evolutivo entre ambas arquitecturas.

Tabla 4: Taxonomía de los procesos de carga operativa.

Carga repetitiva controlada. *(Algoritmos clásicos en CPU)*	Ejecución de bucle continuo de ajuste (*fit()*) y predicción (*predict()*), durante 60 minutos sobre entorno de procesamiento central.
Entrenamiento real por épocas. *(Deep Learning y Modelos Generativos en GPU)*	Reajuste iterativo de pesos mediante retropropagación sobre lotes de datos en 60 minutos de cómputo supervisado.
Inferencia Continua y Sostenida. *(Audio, Agentes, Multimodalidad y arquitectura RAG en GPU)*	Flujo ininterrumpido de procesamiento de secuencias de entrada, decodificación autorregresiva y generación de respuestas durante 60 minutos.

***Nota**:* la carga operativa se ejecutó sobre todos los experimentos bajo las mismas condiciones de entorno.

Para el cálculo de los valores destinados a las métricas, se han agrupado en series de expresiones matemáticas según su naturaleza, organizándolas como: métricas físicas y ambientales que se presentan en la tabla a continuación, también están presentes las métricas de rendimiento técnico – calidad predictiva, y por último las métricas normalizadas de eficiencia energética.

Tabla 5: Expresiones matemáticas para cálculo de métricas físicas y ambientales.

Potencia Media Sostenida (P_{media})	$P_{media} = \frac{1}{T}\int_0^T P(t)\,dt \approx \frac{\sum_{k=1}^{N} P_k}{N}$	P_k: Potencia instantánea en Watts (W) registrada por NVML (GPU) o RAPL/TDP (CPU). N: Muestras en el intervalo T=60 min.
Consumo Energético Total (E_{total})	$E_{total} = \int_0^T P(t)\,dt \approx \sum_{k=1}^{N} P_k \cdot \Delta t_k$	Δt_k: Intervalo discreto de muestreo (10 min). E_{total}: Expresado en Kilovatios-hora (kWh).
Huella de Carbono ($Emisiones_{CO_2eq}$)	$Emisiones_{CO_2eq} = E_{total} \times CI_{red}$	CI_{red}: Factor de intensidad de carbono del combinado eléctrico regional ($_gCO_2eq/kWh$)

***Nota**:* las fórmulas representan los valores calculados al final de cada experimento.

Las expresiones matemáticas correspondientes a la sección de métricas de rendimiento técnico y calidad predictiva no serán detalladas ya que forman parte de los conceptos especializados en estos algoritmos, pero abordan métricas tales como: Error Cuadrático Medio (*MSE*), Raíz de MSE (*RMSE*), Error Absoluto Medio (*MAE*), Exactitud (*Accuracy*), Precisión (*Precision*), Pérdida Logarítmica (*Log Loss*), Índice Davies-Bouldin (*DB*), Intersección sobre Unión (*IoU*), Factor Tiempo Real (*RTF*), Exact Match (*EM*), son ecuaciones que se encuentran detalladas a profundidad en la literatura correspondiente a cada paradigma y que para este estudio pasan a un segundo plano de importancia ya que el enfoque principal es el consumo energético.

A continuación, se detallan las fórmulas para las métricas unitarias normalizadas, las cuales representan indicadores que relacionan la energía consumida, con la unidad fundamental de trabajo útil, estas varían según el dominio algorítmico estudiado y se detallan en la tabla siguiente.

Tabla 6: Expresiones matemáticas para el cálculo de métricas normalizadas.

Carga repetitiva (CPU) Consumo por ciclo de ajuste (E_{fit})	$E_{fit} = \frac{E_{total}}{N_{fits}}$	$\mu Wh/fit$
Carga repetitiva (CPU) Consumo por muestra ($E_{muestra}$)	$E_{muestra} = \frac{E_{total}}{N_{muestras}}$	$\mu Wh/muestra$
Entrenamiento por épocas (GPU) Consumo por época ($E_{época}$)	$E_{época} = \frac{E_{total}}{N_{épocas}}$	$mWh/época$ o $\mu Wh/época$:
LLMs, Agentes, RAG (GPU) Consumo por 1k Tokens Totales ($E_{1k\ tok}$)	$E_{1k_tok} = \left(\frac{E_{total}}{N_{in} + N_{out}}\right) \times 1{,}000$	$Wh/1k\ tokens$
LLMs, Agentes, RAG (GPU) Consumo por Tokens Generados ($E_{1k\ out}$)	$E_{1k_out} = \left(\frac{E_{total}}{N_{out}}\right) \times 1{,}000$	$Wh/1k\ tokens\ out$
Modelo Audio ASR (GPU) (E_{hora_audio})	$E_{hora_audio} = \frac{E_{total} \times 1{,}000}{Horas_{audio}}$	$\mu Wh/hora\ audio$
Sistema RAG SQuAD (GPU) Consumo por consulta RAG ($E_{consulta}$)	$E_{consulta} = \frac{E_{total}}{N_{consultas}}$	$\mu Wh/consulta$

***Nota**:* la variabilidad de los paradigmas requirió distintas fórmulas para definir las unidades de trabajo útil.

Para el factor de medición en el consumo energético, se ha empleado en cada uno de los experimentos el uso de la librería Codecarbon, lanzada oficialmente el 1 de diciembre de 2020, desarrollada

conjuntamente entre el instituto Mila (*Quebec Artificial Intelligence Institute*), BCG GAMMA (*la división global de ciencia de datos y analítica avanzada de Boston Consulting Group*), Haverford College y Comet.ml (*plataforma de gestión de experimentos en MLOps*), con el propósito de proveer software ligero y de código abierto que permita estimar cuantitativamente la energía consumida y la huella de carbono asociada, producida por la ejecución de código tanto en entornos locales como de nube (Schmidt et al., 2021).

Cuenta con 3 modos de rastreo en código, estos son: "*EmissionsTracker*" es el modo estándar con conexión a Internet, "*OfflineEmissionsTracker*" específico para clústeres cerrados, y *"@track_emissions*" para auditar ejecuciones sin modificar el código fuente (Anthony et al., 2020).

Codecarbon integra y se apoya en dependencias a librerías del ecosistema de Python tales como: "*pynvml*" (interfaz de comunicación de bajo nivel), "*psutil y py-cpuinfo*" (detección y monitoreo de CPU y memoria), "*pandas y arrow*" (manipulación de series temporales de emisiones y exportación de archivos), "*requests*" (consultas HTTP para geolocalización de IPs y recuperación de factores de emisión de redes eléctricas regionales), "*click y rapidfuzz*" (construcción de la interfaz de línea de comandos y emparejamiento difuso de nombres de componentes o regiones), "*prometheus_client*" (exportación opcional de métricas de consumo en tiempo real hacia paneles de supervisión).

Con respecto a los conjuntos de datos (*datasets*) empleados para entrenar los modelos, es necesario indicar que, con el firme propósito de garantizar la reproducibilidad científica de los experimentos y prevenir "*cuellos de botella*" por restricciones de acceso, además de eximir la investigación de cualquier conflicto de propiedad intelectual o derechos de autor, todos los conjuntos de datos corresponden a repositorios de acceso abierto (*Open Access*), cuentan con licencias libres permisivas (*MIT, CC BY 4.0, CC BY-SA 4.0, BSD, Public Domain*) o fueron generados mediante mecanismos deterministas con control de parámetros.

Así se detallan para cada paradigma el conjunto de datos empleado: Regresión Lineal (utilizó "*Synthetic Energy*"), KNN (utilizó "*Heart Disease Risk 2026*") y Random Forest (utilizó "*Social Network Ads*"), XGBoost (utilizó "*Synthetic Fraud Dataset*"), K-Means Clustering (utilizó "*Customer Segmentation*"), Autoencoders (utilizó "*Credit Card Fraud Detection*"), Perceptrón Multicapa – MLP (utilizó "*Covertype*"), LSTM Pura NLP (utilizó "*Fake News Detection*"), Fine Tunning de DistilBERT (utilizó

"*IMDB Large Movie Review Dataset*"), ResNet-18 (utilizó "*CIFAR-10 Dataset*"), YOLOv8n (utilizó "*BCCD Blood Cells Count Detection*"), Late Fusion Multimodal (utilizó "*Hateful Memes*"), DDPM Difusión Visual (utilizó "*MNIST Hand-written Digits*"), ASR Faster Whisper (utilizó "*PolyAI/MInDS-14*"), Inferencia Agéntica (utilizó un banco de 10 tareas de razonamiento *ReAct*), Inferencia con Arquitectura RAG (utilizó "*SQuAD v1.1*"). Todos ellos disponibles en los repositorios de: *Scikit-Learn, Kaggle, OpenML, Huggin Face, University of Toronto, Roboflow, Synthetic ReAct Agent* respectivamente.

Resultados.

Llegado a este punto, la fase experimental del proyecto entregó datos tabulables, es importante indicar que se requirió un total de 51 horas de cómputo para ejecutar todos los paradigmas en repetición triple, a excepción del paradigma "*Autoencoder*", el cual se ejecutó 6 repeticiones (3 en CPU, 3 en GPU), y resulta imperativo describir los hallazgos obtenidos clasificándolos por familias de la siguiente manera: *Familia 1 (Aprendizaje Tabular y Clásico).* Estos algoritmos exhibieron un perfil de consumo eléctrico caracterizado por una altísima estabilidad física y baja variabilidad térmica.

Familia 2 (Deep Learning y Procesamiento de Lenguaje Natural). La transición hacia las redes neuronales profundas y la aceleración por hardware dedicado (GPU NVIDIA Tesla T4), incrementó sustancialmente la demanda eléctrica y el consumo acumulado.

Familia 3 (Visión por Computadora y Multimodalidad). En estos paradigmas, las cargas de convolución bidimensional y fusión de modalidades, exigieron un uso intensivo del ancho de banda de VRAM y de los *Tensor Cores* de la tarjeta.

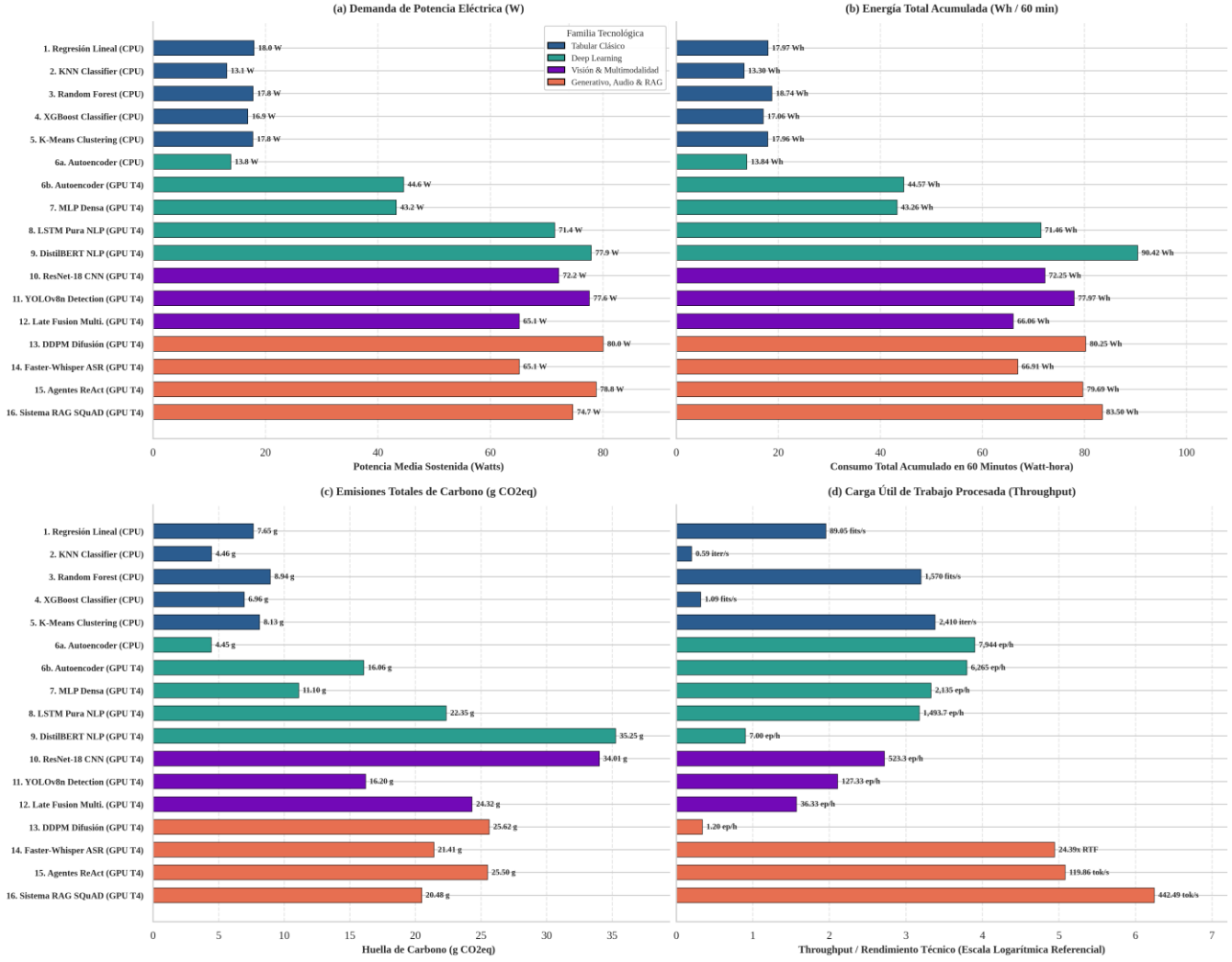
Familia 4 (Modelos Generativos, Audio, Agentes y arquitectura RAG). Las tecnologías más recientes empleadas actualmente, los modelos generativos estocásticos, los decodificadores autorregresivos y las canalizaciones de contexto, representaron las cargas de mayor intensidad eléctrica e impacto ambiental. Para fines ilustrativos, la información completa y tabulada de todos los experimentos, será presentada en la siguiente tabla de contenidos, la cual detalla la información del paradigma evaluado, las métricas relacionadas a la Potencia Media (W), Consumo Total (kWh), Emisiones (gCO_2eq), rendimiento técnico y la métrica unitaria normalizada de los 16 paradigmas, posteriormente se muestran los resultados de manera gráfica de las tres ejecuciones realizadas a cada uno y sintetizadas en torno al modelo.

Tabla 7: Matriz comparativa de los 16 Experimentos.

Paradigma	Potencia Media W	Consumo Total (wKh)	Emisiones (gCO_2eq)	Rendimiento Técnico	Métrica Unitaria Normalizada
Regresión Lineal	17.97 $\pm$ 0.01	0.017970 $\pm$ 0.000003	7.65 $\pm$ 3.42	89.05 *fits/s* $R^2 = 0.9277$	0.0572 μWh / *fit* \$0.00114 μWh /*muestra*
KNN Classifier	13.12 $\pm$ 0.04	0.013304 $\pm$ 0.000075	4.46 $\pm$ 1.35	0.59 *iter/s* $Acc = 98.62\%$	6.2134 μWh / *ciclo*
Random Forest	17.81 $\pm$ 0.01	0.018739 $\pm$ 0.000498	8.94 $\pm$ 3.99	1,570 *fits/s* $Acc = 90.26\%$	35.61 μWh / *fit* 0.00312 μWh / *muestra*
XGBoost Classifier	16.87 $\pm$ 0.09	0.017059 $\pm$ 0.000062	6.96 $\pm$ 5.16	1.09 *fits/s* $Acc = 90.16\%$	4.3047 μWh / *fit* 0.00432 μWh / *muestra*
K-Means Clustering	17.78 $\pm$ 0.03	0.017960 $\pm$ 0.000031	8.13 $\pm$ 3.89	2,410 *iter/s* $s = 0.2398$	0.00198 μWh / *muestra*
Autoencoder CPU	13.83 $\pm$ 0.67	0.013838 $\pm$ 0.000678	4.45 $\pm$ 2.03	7,944 *épocas/h* $AUC = 0.9632$	4.84 μWh / *época*
Autoencoder GPU	44.56 $\pm$ 2.40	0.044574 $\pm$ 0.002408	16.06 $\pm$ 8.64	6,265 *épocas/h* $AUC = 0.9424$	18.45 μWh / *época*
MLP Densa	43.25 $\pm$ 2.45	0.043255 $\pm$ 0.002452	11.10 $\pm$ 4.96	2,135 *épocas/h* $Acc = 88.72\%$	20.27 μWh / *época*
LSTM Pura NLP	71.44 $\pm$ 2.45	0.071456 $\pm$ 0.002443	22.35 $\pm$ 11.68	1,493.7 *épocas/h* $Acc = 97.55\%$	47.84 μWh / *época*
DistilBERT NLP	77.93 $\pm$ 0.26	0.090418 $\pm$ 0.000400	35.25 $\pm$ 1.83	7.00 *épocas/h* $Acc = 88.30\%$	12.92 *mWh* / *época* 12,916.88 μWh / *época*
ResNet-18 CNN	72.15 $\pm$ 0.29	0.072250 $\pm$ 0.000271	34.01 $\pm$ 1.54	523.3 *épocas/h* $Acc = 87.09\%$	138.19 μWh / *época*
YOLOv8n Detection	77.59 $\pm$ 0.17	0.077968 $\pm$ 0.000177	16.20 $\pm$ 5.74	127.33 *épocas/h* $mAP50 = 92.83\%$	612.41 μWh / *época*
Late Fusion Multi.	65.13 $\pm$ 1.40	0.066055 $\pm$ 0.001121	24.32 $\pm$ 11.53	36.33 *épocas/h* $Acc = 62.37\%$	1.82 *mWh* / *época* 1,818.65 μWh / *época*
DDPM Difusión	80.05 $\pm$ 0.45	0.080251 $\pm$ 0.000450	25.62 $\pm$ 11.44	1.20 *épocas/h* $MSE = 0.03736$	66.71 *mWh* / *época* 66,710.0 μWh / *época*
Faster-Whisper ASR	65.13 $\pm$ 2.57	0.066905 $\pm$ 0.000932	21.41 $\pm$ 9.56	24.39x RTF $WER = 12.45\%$	2.7462 *Wh / hora audio* 0.7628 μWh / *s audio*
Agentes ReAct	78.81 $\pm$ 0.18	0.079692 $\pm$ 0.001342	25.50 $\pm$ 11.39	119.86 *tok/s* *Éxito* $= 94.2\%$	0.6604 *Wh / 1k tok out* 47.58 μWh / *paso*
Sistema RAG SQuAD	74.67 $\pm$ 6.25	0.083498 $\pm$ 0.008003	20.48 $\pm$ 4.46	442.49 *tok/s* $Top1 = 65.11\%$	10.53 μWh / *consulta* 0.6837 *Wh / 1k tok out*

***Nota**:* "Rendimiento Técnico" identifica valores de despliegue real en producción para cada algoritmo.

Figura 1. Gráfica consolidada de los resultados obtenidos con los experimentos.



Fuente: elaboración del autor con librería *"matplotlib.pyplot"* de *Python.*

Discusión.

El análisis de la carga energética, a través de los 16 paradigmas revela que la demanda se estratifica en función al hardware y la estructura del algoritmo, los algoritmos clásicos mantuvieron un régimen de bajo consumo, la baja variabilidad observada confirma que, el factor determinante de la potencia es la tasa de ocupación de los hilos de procesamiento, y no la complejidad matemática interna. La GPU mostró dos subniveles de régimen de potencia claramente definidos: las redes densas moderadas, mostraron un estado de carga parcial, donde las transferencias de RAM-VRAM limitan el uso de los *Tensor Cores* y, por otro lado, las Redes Convolucionales, Transformadores y Modelos Generativos, dispararon la demanda energética, alcanzando el límite térmico efectivo del acelerador.

Uno de los hallazgos más trascendentes del estudio es la cuantificación de la penalización energética introducida por la atención en Transformadores (*Transformers*) y los flujos autorregresivos, en comparativa *NLP (LSTM) vs. DistilBERT*, reemplazar la arquitectura recurrente *LSTM* por la atención bidireccional de *DistilBERT* incrementó la potencia media sostenida en 9.1% y elevó el consumo por época masivamente representando un incremento de 269%.

Mientras tanto, al evaluar la generación autorregresiva de texto con el modelo *Qwen1.5-1.8B-Chat* con precisión (*float16*), el consumo energético por 1,000 tokens generados convergió entre la Inferencia Agéntica y la Inferencia con arquitectura RAG, lo cual demuestra que el costo energético de redacción por token es una constante física intrínseca del modelo de lenguaje (*LLM*) en el hardware evaluado. Además a pesar que estos dos últimos paradigmas mencionados utilizan el mismo modelo, la arquitectura RAG resultó ser 4.5 veces más eficiente en términos energéticos por consulta, que un paso de razonamiento agéntico, esto se explica porque la búsqueda vectorial con FAISS (*Facebook AI Similarity Search*) resuelve la recuperación de información en milisegundos, mientras que el ciclo *ReAct* del agente exige múltiples iteraciones completas con autocorrección para completar una tarea.

Un análisis comparativo entre desempeño técnico y consumo eléctrico permite delimitar la frontera de eficiencia conocida como Frontera de Pareto, así encontramos la *"Zona de Alta Eficiencia (IA Verde)",* en la cual algoritmos como XGBoost, Random Forest, ResNet-18 y Faster Whisper en CPU y GPU respectivamente demostraron equilibrio óptimo; del mismo modo se identificó la *"Zona de Penalización Energética (IA Roja)"*, en donde los algoritmos como el modelo de difusión visual DDPM representó el caso extremo de intensidad energética, alcanzando el pico más alto de todo el experimento (66,710.0 μWh), esto se puede justificar con el hecho que la inversión estocástica de 300 pasos de ruido para sintetizar imágenes pequeñas demuestra que los modelos generadores difusos imponen un costo ambiental que debería sopesarse críticamente frente a alternativas no generativas.

Existe también un hallazgo metodológico en el modelo YOLOv8n, en el cual se identificó una particularidad de desacople entre los *callbacks* de entrenamiento (*on_train_epoch_end*) y el rastreador global de CodeCarbon ejecutándose sobre el modelo, al finalizar el ciclo de ejecución de 60 minutos, la función (*model.train()*) de *Ultralytics* ejecuta automáticamente una fase de evaluación post-entrenamiento sobre el conjunto de validación utilizando el método (*best.pt*), debido a que los *callbacks*

dejan de reportar al concluir el bucle de entrenamiento, el medir únicamente mediante escuchadores (*listeners*) omitió 0.061775 *kWh*, que representa un 83.5% del consumo total real de 0.077968 *kWh*, causando una subestimación de la medición real, esto demuestra que la auditoría energética en sistemas de I.A. debe basarse en monitores globales de proceso como por ejemplo (*tracker.stop()*) para evitar estas inconsistencias.

CONCLUSIONES

Se demostró la existencia de dos regímenes térmico-eléctricos (perfil de potencia), diferenciados en la infraestructura evaluada: una demanda contenida y estable en CPU frente a un estado de saturación en GPU T4 para arquitecturas con mecanismos de atención, visión y generación estocástica.

La evaluación del caso *"puente generacional (Autoencoder)"* confirmó que acelerar por GPU una red densa pequeña consume 3.22 veces más energía que la CPU, estableciendo que la aceleración por hardware solo se justifica cuando el volumen de parámetros satura el paralelismo de los núcleos CUDA.

Se identificó que el consumo de generación autorregresiva representa una convergencia energética en generación *LLM* concretamente en el modelo *Qwen1.5-1.8B-Chat* ejecutado en formato de media de precisión de punto flotante de 16 bits (*"float16"* – representación numérica que reduce a la mitad el uso de VRAM respecto al valor nominal *"float32"* y optimiza la ejecución por *Tensor Cores* en GPU T4), constituye una constante de carácter físico; y que el sistema RAG es 4.5 veces más eficiente por consulta que la inferencia agéntica.

Se probó que evaluar la eficiencia energética únicamente mediante *callbacks* de entrenamiento, en arquitecturas *end-to-end* (como YOLOv8) subestima hasta un 83% el consumo real, demostrando la necesidad de emplear rastreadores de procesos a nivel global.

Se estableció una jerarquía de impacto energético absoluto, estableciendo que *Fine Tunning de Transformers (DistilBERT), canalizaciones de conocimiento aumentado RAG, y la síntesis generativa visual DDPM* encabezan el consumo energético acumulado, demostrando que la atención bidireccional y la generación estocástica constituyen la mayor fuente de huella de carbono registrada por este estudio.

REFERENCIAS BIBLIOGRÁFICAS.

Aczel, M., Chamanara, S., Matin, M., Farsi, A., Marwala, T., & Madani, K. (2026). *The Environmental Cost of AI's Energy Use: Carbon, Water and Land Footprints*. https://doi.org/10.53328/inr26rma002

Anthony, L. F. W., Kanding, B., & Selvan, R. (2020). *Carbontracker: Tracking and predicting the carbon footprint of training deep learning models* [Preprint]. arXiv. doi.org

Carneiro, T., Da Nóbrega, R. V. M., Nepomuceno, T., Bian, G.-B., De Albuquerque, V. H. C., & Rebouças Filho, P. P. (2018). *Performance analysis of Google Colaboratory as a tool for accelerating deep learning applications.* IEEE Access, 6, 61677–61685. doi.org

Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R. A., Goodman, S., Townsend, B. C., Ranganathan, P., Dean, J., Vahdat, B., Gomes, B., & Manyika, J. (2025). *Measuring the environmental impact of delivering AI at Google Scale* [Preprint]. arXiv. doi.org

Google. (2026). *Google 2026 Environmental Report*. Google Sustainability. https://sustainability.google/google-2026-environmental-report/

Lin, P., Bunger, R., y Avelar, V. (2023). *Guía de métricas de sostenibilidad medioambiental para centros de datos* (Documento Técnico 67, versión 2). Centro de Investigaciones para la Gestión Energética, Schneider Electric.

Luccioni, A. S., Jernite, Y., & Strubell, E. (2024). Power hungry processing: Watts driving the cost of AI deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 85–99). Association for Computing Machinery. doi.org

Luccioni, A. S., Sylvain, V., Vassilesvski, M., Mitchell, M., & Strubell, E. (2023). Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research*, *24*(253), 1–25.

Marinoni, A., Liò, P., Cambria, E., Dal Zilio, L., Lin, W., Dalla Mura, M., Chanussot, J., Ragusa, E., Mengaldo, G., Tso, C. Y., Zhu, Y., & Horton, B. (2026). *The data heat island effect: quantifying the impact of AI data centers in a warming world*. arXiv. doi.org

Mistral AI. (22 de julio de 2025). *Our contribution to a global environmental standard for AI*. Mistral AI Blog. mistral.ai

Rosa, L., Sangiorgio, M., Scanlon, B. R., Mazzotti, M., Berry, J., & Michalak, A. M. (2026). *One quarter of global data centers are located in water-stressed areas. Environmental Science & Technology*, 60(39), 27780–27791. https://doi.org/10.1021/acs.est.6c08088

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

Schmidt, V., Goyal, K., Joshi, A., Victor, V., Luccioni, S., O'Connor, M., & Lakhani, N. (2021). *CodeCarbon: Tracking carbon emissions from machine learning code* [Software Informático]. Zenodo. https://doi.org/10.5281/zenodo.4658424

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645–3650). Association for Computational Linguistics. doi.org

Tangermann, V. (9 de julio de 2026). *The Pollution Being Churned Out by AI Data Centers Is So Severe That It's Almost Incomprehensible*. Futurism. https://futurism.com/artificial-intelligence/pollution-ai-data-centers-severe