

PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE CIENCIAS EXACTAS NATURALES Y AMBIENTALES

CARRERA DE CIENCIA DE DATOS

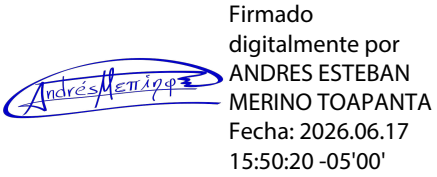
Estado del arte de los Modelos de Lenguaje de Gran Escala: arquitecturas, entrenamiento y aplicaciones actuales

Monografía previa a la obtención del título de Ingeniero en Ciencia de Datos

Martín Baca Sandoval

Quito, 2026

Certifico que la Monografía de Ciencia de Datos de **Martín Baca Sandoval** ha sido concluida de conformidad con las normas establecidas; por lo tanto, puede ser presentada para la calificación correspondiente.



Andrés Esteban Merino Toapanta

Tutor de la monografía

Quito, 25 de junio de 2026

DEDICATORIA

A mis papas, mi hermano y mi sobrina, quienes me apoyaron incondicionalmente durante el transcurso de mi carrera.

AGRADECIMIENTOS

A mis papas, Germán y Josseth, por todo el esfuerzo que pusieron para que yo pueda finalizar la carrera. Gracias a su esfuerzo he podido salir adelante en todos los aspectos de mi vida y con esa misma fuerza continuaré mi camino.

A mi hermano, Gabriel, por su apoyo constante y por ser mi compañero de camino durante mi formación académica. Su presencia y palabras de aliento fueron fundamentales para mantener la motivación en los momentos más difíciles.

A mi querido amigo y socio, Kevin Rojas, gracias por introducirme en el mundo de la inteligencia artificial y por acompañarme en el camino de emprender. Valoro mucho tu apoyo constante durante el desarrollo de esta investigación.

A mi tutor, Andrés Merino, por haberme impulsado desde el día 1 de la carrera y por apoyarme con la elaboración de mi monografía.

Al equipo dinamita, Sofía y Melani, por hacer la carrera más agradable y por compartir tantos proyectos, risas y trabajos en conjunto.

TABLA DE CONTENIDOS

1. Resumen **1**

2. Abstract **2**

3. Introducción **3**

4. Objetivos **7**

4.1. Objetivo general 7

4.2. Objetivos específicos 7

5. Fundamento teórico **8**

5.1. Fundamentos de los Modelos de Lenguaje de Gran Escala (LLMs) 8

5.1.1. Evolución del Procesamiento de Lenguaje Natural hacia los LLMs 8

5.1.2. El paradigma de los Modelos Fundacionales y las leyes de escalado 9

5.1.3. La arquitectura base: El Transformer Autoregresivo 12

5.2. El Mecanismo de Atención Estándar 13

5.2.1. Funcionamiento matemático de la Atención de Producto Punto Escalado . 13

5.2.2. El rol del KV Cache en la generación de texto 15

5.3. Limitaciones Computacionales de la Atención 17

5.3.1. La barrera de la complejidad cuadrática ($\mathcal{O}(n^2)$) en tiempo y memoria . . 17

5.3.2. El cuello de botella en la memoria de la GPU (VRAM) durante la inferencia 18

5.3.3. Desafíos al procesar ventanas de contexto largo: degradación de rendimiento y el fenómeno *Lost in the Middle* 20

5.4. Estrategias arquitectónicas alternativas al mecanismo de atención estándar 22

5.4.1. Atención Dispersa (Sparse Attention) 22

5.4.2. Atención Lineal (Linear Attention) 23

5.4.3. Modelos de Espacio de Estados (SSM) 23

5.5. Fundamentos de Métricas de Evaluación en LLMs 24

5.5.1. Métricas de Eficiencia Computacional: Definición de FLOPs, latencia, rendimiento y consumo energético 24

5.5.2. Métricas de Calidad y Benchmarks: Qué miden la Perplejidad y los benchmarks estandarizados 27

6. Metodología **31**

6.1. Protocolo de revisión 31

6.2. Preguntas de investigación 31

6.3. Descripción de la estrategia de búsqueda 33

6.4. Selección de estudios 33

6.5. Selección de estudios 33

6.6. Síntesis de la literatura y análisis 34

7. Resultados y discusión **36**

7.1. Perfil de los estudios incluidos 36

7.2. RQ1: ¿Qué arquitecturas y componentes aparecen con mayor frecuencia en LLMs entre 2020–2025 y cómo evoluciona su proporción por año? 36

7.2.1. Sparse Attention (Atención Dispersa) 37

7.2.2. Linear Attention (Atención Lineal) 39

7.2.3. State Space Models (SSM) 40

7.2.4. Mezcla de Expertos (MoE) 40

7.2.5. Arquitecturas Híbridas 41

7.3. RQ2: ¿Qué rangos y medianas reportan las principales estrategias arquitectónicas en eficiencia y en calidad durante 2020–2025 y cuál es la diferencia mediana entre estrategias comparables bajo condiciones declaradas? 42

7.3.1. Latencia e Inferencia (Throughput) 43

7.3.2. Uso de Memoria y VRAM 44

7.3.3. Complejidad Computacional (FLOPs) 45

7.3.4. Calidad Predictiva y Benchmarks de Rendimiento 46

7.3.5. Eficiencia Energética y Huella de Carbono 47

7.4. RQ3: ¿Qué proporción de estudios reporta mejoras frente a limitaciones de atención y de qué magnitud por tipo de enfoque? 48

7.4.1. La Era de la Consolidación (2020–2022): Hegemonía de la Atención Cuadrática 50

7.4.2. El Punto de Inflexión (2023): La Transición Sub-Cuadrática 50

7.4.3. El Cambio de Paradigma (2024): Consolidación Experimental de las Arquitecturas Híbridas 51

7.4.4. Tendencias Consolidadas y Perspectivas (2025) 51

8. Conclusiones **54**

9. Referencias **56**

1. RESUMEN

Los modelos de lenguaje de gran escala (LLM) han experimentado un rápido desarrollo desde la introducción del mecanismo de atención transformador, cuya complejidad cuadrática respecto a la longitud de secuencia constituye su principal cuello de botella. Este trabajo presenta una revisión sistemática de la literatura, siguiendo el protocolo PRISMA 2020, orientada a caracterizar el estado del arte de las arquitecturas alternativas o híbridas propuestas entre 2020 y 2025 para superar dicha limitación. A partir de una búsqueda en Scopus que arrojó 359 registros, se seleccionaron 31 artículos mediante criterios de inclusión y exclusión basados en el marco PICOC. Se abordaron tres preguntas de investigación: clasificación arquitectónica (RQ1), métricas de eficiencia y calidad reportadas (RQ2) e impacto de las mejoras propuestas (RQ3). Los resultados muestran que las arquitecturas híbridas dominan el corpus (18/31, 58,1 %), seguidas de Sparse Attention y Mixture of Experts (6/31 cada una), con un único estudio basado en State Space Models y ninguno de Linear Attention pura. La memoria fue la métrica más cubierta (25/31), con una mediana de 8,0 GB para modelos híbridos. El 58,1 % de los estudios aborda explícitamente las limitaciones de la atención estándar; el 93,5 % reporta mejoras de eficiencia y el 90,3 % mejoras de calidad, sin evidencia de un compromiso sistemático entre ambas dimensiones. Se concluye que la comunidad investigadora transita de manera pragmática hacia soluciones híbridas que preservan la capacidad expresiva del transformer reduciendo su costo computacional.

Palabras clave: modelos de lenguaje de gran escala, Sparse Attention, Mixture of Experts, State Space Models, revisión sistemática, PRISMA 2020.

2. ABSTRACT

Large language models (LLMs) have undergone rapid development since the introduction of the transformer attention mechanism, whose quadratic complexity with respect to sequence length represents its primary computational bottleneck. This paper presents a systematic literature review, following the PRISMA 2020 protocol, aimed at characterising the state of the art of alternative and hybrid architectures proposed between 2020 and 2025 to overcome this limitation. Starting from a Scopus search that retrieved 359 records, 31 articles were selected through inclusion and exclusion criteria based on the PICOC framework. Three research questions were addressed: architectural classification (RQ1), reported efficiency and quality metrics (RQ2), and the impact of the proposed improvements (RQ3). Results show that hybrid architectures dominate the corpus (18/31, 58.1 %), followed by Sparse Attention and Mixture of Experts (6/31 each), with a single State Space Model study and no instance of pure Linear Attention. Memory was the most frequently reported metric (25/31), with a median of 8.0 GB for hybrid models. Some 58.1 % of studies explicitly address limitations of standard attention; 93.5 % report efficiency improvements and 90.3 % quality improvements, with no evidence of a systematic trade-off between the two dimensions. The findings suggest that the research community is pragmatically transitioning towards hybrid solutions that preserve the expressive capacity of the transformer while reducing its computational cost.

Keywords: large language models, Sparse Attention, Mixture of Experts, State Space Models, systematic review, PRISMA 2020.

3. INTRODUCCIÓN

El Procesamiento de Lenguaje Natural (PLN) ha atravesado una transformación sin precedentes en la última década, evolucionando desde sistemas basados en reglas y modelos estadísticos tradicionales hacia arquitecturas de aprendizaje profundo masivo (Goodfellow et al., 2016). El punto de inflexión fundamental para esta disciplina ocurrió en 2017 con la introducción de la arquitectura *Transformer* por Vaswani et al. (2017). Al descartar por completo las dependencias secuenciales de las redes neuronales recurrentes y depender exclusivamente del mecanismo de atención, esta arquitectura permitió una paralelización computacional masiva (Vaswani et al., 2017).

Este avance sentó las bases para el surgimiento de los Modelos de Lenguaje de Gran Escala, por sus siglas en inglés LLMs, los cuales han demostrado capacidades emergentes notables para la comprensión, generación y razonamiento sobre el lenguaje humano (Bommasani et al., 2021). Estos sistemas no solo han redefinido la interacción hombre-máquina, sino que han impulsado una nueva frontera de inteligencia artificial capaz de asimilar vastas cantidades de información no estructurada, transformando sectores enteros desde la ingeniería de software hasta la investigación académica.

Durante el periodo comprendido entre 2020 y 2025, el ecosistema de investigación y desarrollo de LLMs ha estado dominado casi de forma absoluta por la variante arquitectónica de «Solo Decodificador» (*Decoder-only*) autorregresivo (Kaplan et al., 2020). Al entrenar estas redes masivas con el objetivo autosupervisado de predecir el siguiente token sobre terabytes de datos, modelos con miles de millones de parámetros han logrado un rendimiento que rivaliza con el humano en una amplia gama de dominios (Brown et al., 2020; Kaplan et al., 2020).

Para cuantificar estas capacidades heurísticas de alta complejidad, la academia ha establecido rigurosos estándares de evaluación (*benchmarks*), exigiendo a los modelos demostrar no solo fluidez lingüística y baja perplejidad, sino también un profundo conocimiento enciclopédico general a través del MMLU (Hendrycks et al., 2021), capacidades de resolución matemática deductiva con pruebas como GSM8K (Cobbe et al., 2021) y la abstracción de razonamiento

científico complejo mediante el *AI2 Reasoning Challenge* (ARC) (Clark et al., 2018).

Sin embargo, el éxito hegemónico del Transformer Decoder-only está intrínsecamente ligado a su mayor debilidad estructural: las profundas limitaciones computacionales impuestas por el mecanismo de atención estándar (*Scaled Dot-Product Attention*) (Vaswani et al., 2017). Matemáticamente, la autoatención exige que cada unidad léxica o token calcule su grado de alineación semántica con todos los demás tokens presentes en el contexto de entrada. Esto se traduce en un cuello de botella algorítmico caracterizado por una complejidad temporal y espacial cuadrática, denotada asintóticamente como $\mathcal{O}(n^2)$ (Dao et al., 2022; Vaswani et al., 2017). A medida que las aplicaciones prácticas exigen a los modelos ingerir secuencias de contexto cada vez más largas, como el análisis de historiales médicos o repositorios de código, esta barrera cuadrática devora exponencialmente los ciclos de las Unidades de Procesamiento Gráfico (GPU) y satura el ancho de banda de la memoria (Dao et al., 2022).

Para mitigar este colapso durante la fase iterativa de generación de texto, la industria ha estandarizado el uso del caché KV (Caché de Clave-Valor) (Pope et al., 2023). Este método almacena los cálculos previos de la matriz de atención para evitar el recálculo, transformando la complejidad temporal de la inferencia en una progresión lineal por cada token generado (Pope et al., 2023). No obstante, esta solución algorítmica engendra un nuevo problema de hardware: un severo cuello de botella en la memoria de la GPU (*VRAM*).

El alojamiento dinámico de los tensores del caché KV satura el ancho de banda de la memoria (*Memory-Bandwidth Bound*), lo que degrada el rendimiento interactivo del modelo al aumentar la latencia del primer token (*Time To First Token* o TTFT) y reducir la tasa de generación continua (Tokens per Second) (Bari et al., 2025; Pope et al., 2023). El impacto de esta ineficiencia trasciende el ámbito del software, derivando en un consumo energético y una huella de carbono elevados para el despliegue a gran escala de la inteligencia artificial generativa (Luccioni et al., 2023; Patterson et al., 2021).

Más allá de los obstáculos de infraestructura y hardware, el intento de procesar ventanas de contexto kilométricas ha revelado profundas falencias en la integridad heurística de la atención estándar. Investigaciones recientes han tipificado el fenómeno conocido como *Lost in the Middle*

(Perdido en el Medio), documentando cómo, incluso en modelos comercializados explícitamente para el manejo de largos contextos, la capacidad de recuperar información pertinente sufre una fuerte degradación cuando los datos críticos se encuentran sepultados en las posiciones intermedias del documento original (N. Liu et al., 2024). En este escenario, la red conserva precisión únicamente en los extremos de la secuencia debido a la prevalencia de los sesgos de primacía y recencia, lo que cuestiona la fiabilidad de las arquitecturas Transformer para el análisis exhaustivo de grandes volúmenes de datos no estructurados (N. Liu et al., 2024).

En este escenario, el problema central radica en que, a pesar de la proliferación de arquitecturas alternativas al Transformer clásico publicadas entre 2020 y 2025, no existe una síntesis estructurada y comparable que permita determinar empíricamente cuáles de estas propuestas logran superar las limitaciones de la atención estándar, la complejidad $\mathcal{O}$, el cuello de botella en VRAM y la degradación del contexto largo, sin comprometer la calidad cognitiva del modelo.

Es precisamente en la intersección de estas crisis operativas, la ineficiencia asintótica $\mathcal{O}(n^2)$, el hambre insaciable de memoria VRAM y la miopía del contexto extenso, donde ha surgido un prolífico y fragmentado campo de investigación arquitectónica. Buscando superar la barrera del Transformer base, la literatura técnica de los últimos cinco años ha propuesto una avalancha de intervenciones. Entre estas destacan los mecanismos de atención dispersa (Sparse Attention), atención lineal, el enrutamiento de sub-redes mediante Mezcla de Expertos (*Mixture of Experts* o MoE) (Shazeer et al., 2017) y más recientemente, el resurgimiento de arquitecturas secuenciales avanzadas como los Modelos de Espacio de Estados (*State Space Models* o SSM), ejemplificados por el modelo Mamba (Gu y Dao, 2023).

La relevancia de esta investigación radica en que la creciente demanda de aplicaciones críticas como el análisis de historiales médicos, la auditoría de contratos legales y el procesamiento de repositorios de código hace urgente identificar qué estrategias arquitectónicas son realmente viables. Una síntesis rigurosa y reproducible bajo lineamientos PRISMA representa un aporte metodológico concreto para investigadores y desarrolladores que necesitan tomar decisiones informadas sobre qué arquitectura adoptar según sus restricciones de hardware y requisitos de calidad.

A pesar de la rápida publicación de estos enfoques prometedores en entornos académicos e industriales, existe una carencia significativa de síntesis secundarias estructuradas. Actualmente, resulta complejo para los investigadores y desarrolladores discernir qué proporción de estas nuevas arquitecturas realmente logra superar el rendimiento de las líneas base (*baselines*) de Solo Decodificador en condiciones experimentales comparables y cuáles son los verdaderos compromisos (*trade-offs*) empíricos entre el ahorro de costes y la retención de la calidad cognitiva (Kitchenham y Charters, 2007).

Para abordar esta brecha en el conocimiento, la presente monografía tiene como propósito realizar una revisión sistemática, estructurada bajo los lineamientos metodológicos de la declaración PRISMA (*Preferred Reporting Items for Systematic reviews and Meta-Analyses*) para asegurar su transparencia y reproducibilidad. El objetivo general es analizar la evolución de los modelos de lenguaje de propósito general (superiores a mil millones de parámetros) y la eficiencia de sus arquitecturas cuando implementan estrategias más allá del mecanismo de atención estándar, evaluando su impacto tanto en la infraestructura de hardware como en la integridad del razonamiento heurístico (Bengio et al., 2021; Kitchenham y Charters, 2007).

4. OBJETIVOS

En este capítulo se presentan los objetivos que orientan la presente revisión sistemática. Se formula primero un objetivo general que delimita el alcance global del estudio y, posteriormente, se desglosan tres objetivos específicos que se corresponden directamente con las preguntas de investigación planteadas en la sección metodológica y articulan los criterios de análisis aplicados al corpus seleccionado.

4.1 OBJETIVO GENERAL

Analizar la evolución de las estrategias arquitectónicas en Large Language Models durante el periodo 2020-2025, caracterizando la frecuencia de sus componentes, el balance eficiencia-calidad reportado y las mejoras frente a las limitaciones de atención.

4.2 OBJETIVOS ESPECÍFICOS

A partir del objetivo general, se derivan tres objetivos específicos que descomponen el análisis en dimensiones operativas y verificables. Cada uno de ellos se aborda mediante una pregunta de investigación independiente y guía la extracción, síntesis y discusión de los resultados presentados en los capítulos posteriores.

- Determinar las arquitecturas y componentes que predominan en los Large Language Models desarrollados entre 2020 y 2025.
- Comparar los rangos y medianas de eficiencia y calidad reportados por las principales estrategias arquitectónicas.
- Evaluar la proporción y magnitud de las mejoras reportadas en estudios que abordan limitaciones de atención.

5. FUNDAMENTO TEÓRICO

5.1 FUNDAMENTOS DE LOS MODELOS DE LENGUAJE DE GRAN ESCALA (LLMS)

Para comprender la evolución y las limitaciones de las arquitecturas actuales, es necesario revisar los fundamentos sobre los que se construyen los Modelos de Lenguaje de Gran Escala, desde sus orígenes en el Procesamiento de Lenguaje Natural hasta los paradigmas de escalado que definen el estado del arte actual.

5.1.1 EVOLUCIÓN DEL PROCESAMIENTO DE LENGUAJE NATURAL HACIA LOS LLMS

El Procesamiento de Lenguaje Natural (PNL) es una de las subramas de la Inteligencia Artificial que se especializa en entender, interpretar y hasta generar lenguaje y ha atravesado una serie de cambios paradigmáticos que revolucionarán al mundo este siglo XXI. En sus inicios, los algoritmos de PNL funcionaban únicamente siguiendo un conjunto de reglas, pero luego se comenzaron a utilizar algoritmos de aprendizaje supervisado tales como regresiones logísticas, máquinas de soporte vectorial, modelos ocultos de Markov, entre otros (Giraldo Forero y Orozco Duque, 2023). Sin embargo, estos modelos tienen algunos problemas latentes, como la incapacidad de abstraer la semántica del lenguaje continuamente o comprender relaciones léxicas más profundas.

En 2013 se introdujeron los *Word Embeddings* marcando un primer hito para la representación computacional del lenguaje. Con los modelos Word2Vec que consiguen transformar las palabras a vectores en un espacio vectorial multidimensional . En dicho espacio vectorial las distancias geométricas entre los vectores representan su similitud tanto semántica como sintáctica (Giraldo Forero y Orozco Duque, 2023). De igual forma, también se pasaron de los modelos de aprendizaje supervisados mencionados anteriormente a arquitecturas con redes neuronales recurrentes, como la arquitectura *Long-Short Term Memory* (LSTM) que posibilitan el análisis secuencial de las palabras actualizando un estado oculto por cada palabra procesada (Vaswani et al., 2017). No obstante, este tipo de arquitecturas tienen un problema inherente y es su inca-

pacidad de paralelización a la hora de entrenar en unidades de procesamiento gráfico (GPUs). En 2017 se publicó el artículo que sentaría las bases para la siguiente revolución en el campo de la Inteligencia Artificial: Grandes Modelos de Lenguaje (LLMs). El artículo *Attention Is All You Need* desarrollado por Vaswani et al. (2017) plantea un nuevo tipo de arquitectura denominada *Transformer* (véase sección 5.1.3), con la que las redes neuronales recurrentes se ven reemplazadas por el mecanismo de autoatención (*Self-Attention*). Se define token como la unidad mínima de lenguaje procesada por un modelo; la autoatención permite que cada token en una secuencia interactúe y ponga atención a todos los demás tokens al mismo tiempo, sin importar su posición dentro del texto. La nueva arquitectura no únicamente logró la paralelización durante el entrenamiento en hardware de computación de alto rendimiento, sino que posibilitó la retención de memoria a largo plazo. La Figura 5.1 ilustra esquemáticamente los componentes principales de la arquitectura Transformer, incluyendo los bloques de codificador y decodificador, las capas de atención multi-cabeza y las conexiones residuales con normalización por capa.

Entre 2020 y 2025, el campo del PLN adoptó como un estándar al Transformer, abandonando las arquitecturas anteriores. Después de este punto, comenzó la era del escalado, donde el peso de los textos de entrenamiento sobrepasan los petabytes, además del aumento exponencial del número de parámetros en los modelos, pasando de unos pocos millones a miles de millones y más recientemente hasta billones. Estos incrementos tan sustanciales han logrado la emergencia de capacidades lógicas y de razonamiento para los LLMs más nuevos. La Inteligencia Artificial generativa está logrando reproducir textos que igualan las capacidades humanas en tareas de comprensión lectora, razonamiento matemático y generación de código (Cobbe et al., 2021; Hendrycks et al., 2021).

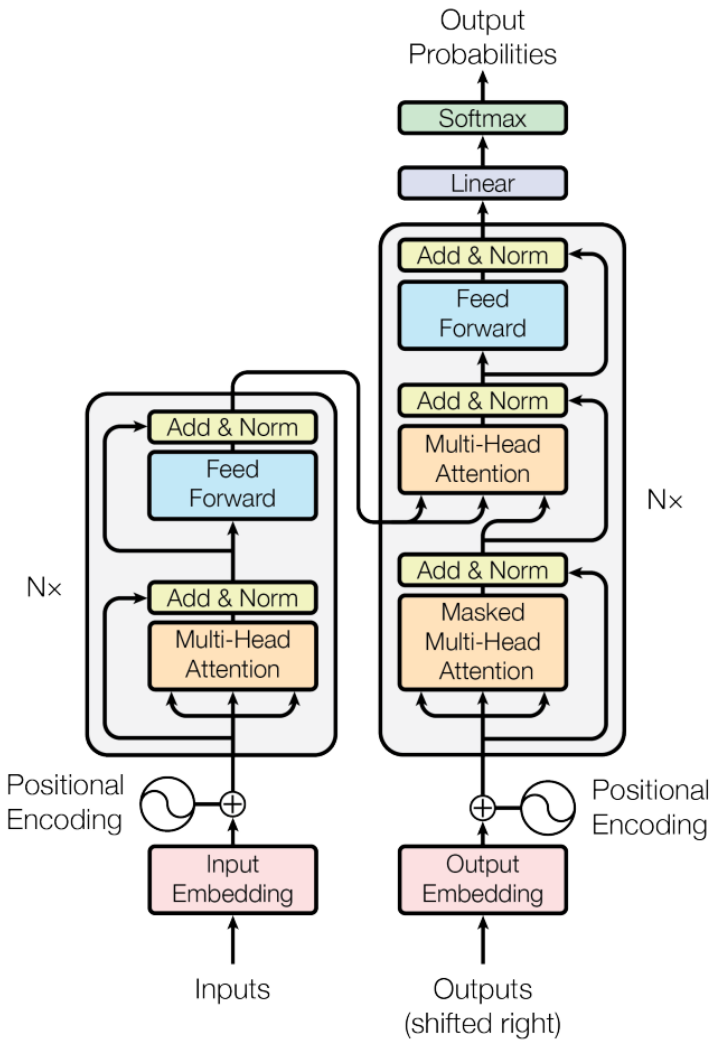
5.1.2 EL PARADIGMA DE LOS MODELOS FUNDACIONALES Y LAS LEYES DE ESCALADO

La introducción de la arquitectura Transformer junto con la computación impulsada por GPUs modernas y corpus de texto masivos han logrado el surgimiento de un nuevo paradigma en el campo de los Modelos Fundacionales. Modelo Fundacional se define como una red neuronal masiva pre-entrenada inicialmente con aprendizaje autosupervisado sobre datos vastos con el fin de alcanzar una generalidad cognitiva que más adelante será adaptada a para resolver tareas más

específicas utilizando técnicas de ajuste fino (*fine-tuning*) (Bommasani et al., 2021). La eficiencia y capacidad de estos modelos dependen de las denominadas Leyes de Escalado (*Scaling Laws*).

Figura 5.1

Diagrama de la arquitectura Transformer



Nota. Tomado de «Attention Is All You Need», por A. Vaswani et al., 2017, *Advances in Neural Information Processing Systems 30*, p. 3.

Previamente se creía que el rendimiento de los modelos se basaba principalmente en la arquitectura que se elija. Sin embargo, Kaplan et al. (2020) determinó empíricamente que las Scaling Laws que proponen que el rendimiento de los modelos construidos sobre la arquitectura Transformer se basa principalmente en 3 factores: Número de parámetros N, el tamaño del conjunto de entrenamiento D y la cantidad de cómputo usada para el entrenamiento C.

Kaplan et al. (2020) propone que mientras no hayan cuellos de botella en ninguna de

las variables, el rendimiento de un modelo mejora de forma predecible. De igual forma concluyó que cuando se tiene una cantidad de cómputo finita determinada, se prioriza el número de parámetros (N) sobre la cantidad de datos (D). Inclusive se llegó a una relación óptima entre las variables, por un lado $N_{opt} \propto C^{0,73}$, mientras que los datos escalan apenas a $D_{opt} \propto C^{0,27}$ (Kaplan et al., 2020). En *Language Models are Few-Shot Learners* de Brown et al. (2020) se concluye que el ratio empirico de entrenamiento es extremadamente bajo ya que son 1.7 tokens por parámetro, deducido del modelo GPT-3 que cuenta con 175 mil millones de parámetros y fue entrenado con 300 millones de tokens. No obstante, Hoffmann et al. (2022) corrige este consenso recopilando los datos de entrenamiento de alrededor de 400 modelos con parámetros que van desde los 70 millones hasta 16 mil millones en 5 a 500 mil millones de tokens. Este estudio encontró que los modelos estaban subentrenados ya que estaban siguiendo lo planteado por Kaplan et al. (2020). Hoffman encontró que en el artículo de Kaplan no se cuentan los parámetros de incrustación (*embedding parameters*. Esto condujo a Hoffman a encontrar nuevas relaciones óptimas, $N_{opt} \propto C^{0,50}$ para el número de parámetros y $D_{opt} \propto C^{0,50}$ para la cantidad de datos (Hoffmann et al., 2022). Además, el artículo *Reconciling Kaplan and Chinchilla Scaling Laws* de Pearce y Song (2024) se confirmó empiricamente que el número de tokens óptimos por parámetro son 20. Según esta corrección GPT-3 debió haberse entrenado con 3.5 billones de tokens para optimizar la arquitectura del modelo.

Por otro lado, en 2024 se halló que para modelos más recientes, particularmente *Llama 3* (AI@Meta et al., 2024) que si se siguen los consensos descubiertos por Hoffman, los modelos más pequeños quedarían subentrenados. El modelo Llama 3 de 8 mil millones de parámetros se entrenó con un corpus de texto de 1.5 billones de tokens, descubriendo una nueva relación de 1,875 tokens por parámetro sin observar algún tipo de degradación en el rendimiento del modelo. El mismo año se hizo un descubrimiento sin precedentes a la hora de usar la arquitectura denominada como *Mixture of Experts* (Krajewski et al., 2024): la relación entre tokens y número de parámetros disminuye a medida que aumenta el número de parámetros, al tener un número fijo de expertos (subredes especializadas) (Krajewski et al., 2024). Se llegó a obtener una relación de 44 tokens por parámetro. Esto último incentiva a la creación de modelos más reducidos pero entrenados con datos que sobre pasan los exabytes de información.

5.1.3 LA ARQUITECTURA BASE: EL TRANSFORMER AUTOREGRESIVO

El esquema inicial del Transformer planteado por Vaswani et al. (2017) muestran una arquitectura conocida como Codificador-Decodificador (*Encoder-Decoder*) que fue concebida específicamente para traducción de lenguaje, de inglés a alemán y de inglés a francés. Esta arquitectura está constituída por 6 capas de codificación y 6 de decodificación, además permite que los codificadores apliquen atención bidireccional no enmascarada con el fin de sintetizar una representación del texto contextual densa; por otro lado, los decodificadores utilizan atención enmascarada sobre sí mismos y atención cruzada hacia las representaciones del codificador para generar la traducción de forma iterativa.

Poco tiempo después de la introducción del Transformer, la comunidad académica estableció 3 paradigmas estructurales primarios, cada uno explotando las eficiencias específicas:

- **Arquitecturas de Solo Codificador (*Encoder-only*):** Representadas por la familia BERT (Devlin et al., 2019), utilizan autoatención bidireccional pura examinando todos los tokens simultáneamente. Destacan en clasificación de secuencias, extracción de información y Reconocimiento de Entidades Nombradas (NER), aunque carecen de capacidad nativa de generación de texto libre.
- **Arquitecturas Codificador-Decodificador (*Encoder-Decoder*):** Conservada en modelos como T5 y BART, es ideal para tareas secuencia a secuencia (seq2seq) como resúmenes y traducción (Nair et al., 2024).
- **Arquitecturas de Solo Decodificador (*Decoder-only*):** Paradigma actual de la IA generativa, representado por GPT. Trata el prompt y la salida como una única secuencia, facilitando una paralelización extrema y demostrando que predecir el siguiente token a escala masiva es suficiente para desarrollar razonamiento general.

.

5.2 EL MECANISMO DE ATENCIÓN ESTÁNDAR

5.2.1 FUNCIONAMIENTO MATEMÁTICO DE LA ATENCIÓN DE PRODUCTO PUNTO ESCALADO

En *Attention Is All You Need* se plantea y formaliza el motor para el funcionamiento de todos los LLMs, el mecanismo de Atención de Producto Punto Escalado (*Scaled Dot Product Attention*). A diferencia de las recurrencias que tienen la desventaja de la señal a lo largo del avance de pasos secuenciales, este mecanismo hace que la red neuronal calcule analíticamente la afinidad semántica y sintáctica exacta entre cualquier par de tokens en una secuencia arbitrariamente larga de manera estrictamente paralela.

El flujo de operaciones inicia cuando cada token de la secuencia de entrada es mapeada a un vector de incrustación de dimensión d_{model}. Luego, con el fin de integrar este token al mecanismo de atención, el vector de entrada se proyecta simultáneamente en tres subespacios latentes distintos mediante una multiplicación por tres matrices de pesos entrenables y parametrizadas de manera independiente: $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$. El resultado de esta operación son 3 matrices:

- Matriz de consultas (*Queries* - Q): Representa el estado informativo actual del token focal. Modela fenomenológicamente qué atributos semánticos o sintácticos está buscando activamente este token en su contexto.
- Matriz de Claves (*Keys* - K): Representa la identidad abstracta de todos los tokens del contexto. Actúa como el indexador de una base de datos, respondiendo a la pregunta «qué características semánticas posee intrínsecamente este token para ofrecer al resto de la red».
- Matriz de Valores (*Values* - V): Contiene la señal informacional latente real que será extraída, ponderada y propagada hacia las capas superiores de la red neuronal si se detecta una alineación exitosa entre la Consulta y la Clave.

El núcleo del algoritmo está en calcular cuánta compatibilidad y alineación hay entre la

Consulta actual y todas las Claves disponibles en el contexto. Esto se consigue al realizar la operación de producto punto. Matemáticamente, el producto escalar entre dos vectores euclidianos captura su similitud de coseno, magnitud latente y alineación direccional en el hiperespacio representacional. Geométricamente, vectores que apuntan en direcciones paralelas, que tienen una alta correlación semántica, producirán un producto escalar fuertemente positivo, mientras que vectores ortogonales, que son conceptos no relacionados, tenderán a cero. Cuando se multiplican las matrices Q por K^T, el modelo computa una matriz cuadrada de puntajes de atención brutos, que hace una relación completa de cada token con respecto a todo su vecindario lingüístico.

A medida que la dimensionalidad de las representaciones latentes (la profundidad de las claves) aumenta para albergar conceptos más complejos, la varianza matemática del producto punto se dispara proporcionalmente. Los productos escalares resultantes adquieren magnitudes algebraicas enormes. Al ingresar estos valores desproporcionados a la función subsiguiente, el modelo sufriría el fenómeno del desvanecimiento de gradiente (*vanishing gradient problem*): las derivadas parciales colapsarían asintóticamente a cero, imposibilitando la retropropagación de errores y deteniendo el aprendizaje de los tensores (Goodfellow et al., 2016).

Para neutralizar esta anomalía, se implementó un paso crítico de escalado: cada elemento de la matriz de puntajes crudos se divide entre un factor de estabilización equivalente a la raíz cuadrada de la dimensión de las claves ($\sqrt{d_k}$) (Vaswani et al., 2017). Esta atenuación matemática asegura que la distribución de los productos mantenga una varianza estable, independientemente de la dimensionalidad del modelo.

La formulación algorítmica unificada se define finalmente mediante la ecuación de la atención escalada:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

Sobre la matriz atenuada se aplica una función de activación no lineal softmax a lo largo del eje dimensional de las secuencias. La función softmax transforma los logaritmos crudos acotándolos en una rigurosa distribución probabilística. Esta distribución actúa como un mecanismo de atención biológica: asigna ponderaciones fuertemente cercanas a 1 a las características contextualmente críticas e ignora estáticamente (con pesos asintóticos a 0) la redundancia sintáctica.

Finalmente, estos pesos probabilísticos normalizados se multiplican linealmente por la matriz de Valores (V). El vector de salida sintetiza una representación ricamente contextualizada, donde la identidad original de cada token se ha enriquecido fusionando la carga semántica de sus vecinos más relevantes (Vaswani et al., 2017).

Este procedimiento matemático no se ejecuta de forma monobloque. La arquitectura moderna utiliza un enfoque de Autoatención de Múltiples Cabezales (*Multi-Head Attention*), dividiendo la dimensionalidad del modelo para ejecutar múltiples proyecciones de en paralelo (Vaswani et al., 2017). Esto faculta a la inteligencia artificial para atender de forma concurrente a información alojada en subespacios representacionales completamente heterogéneos, permitiéndole a una «cabeza» especializarse en resolver anáforas y referencias de pronombres, mientras otra modela rigurosamente el sentimiento pragmático global del párrafo.

5.2.2 EL ROL DEL KV CACHE EN LA GENERACIÓN DE TEXTO

Durante la fase de inferencia y despliegue en la nube, la generación de texto por parte de un Transformer autoregresivo opera bajo un régimen algorítmico iterativo estricto: el modelo ingiere el historial completo de la conversación y emite predictivamente la distribución probabilística para un único token nuevo en cada paso de tiempo discreto t. Inmediatamente, este token recién concebido se concatena al final de la secuencia causal, conformando el nuevo contexto para proyectar el token en el paso $t + 1$ (Vaswani et al., 2017).

Bajo la implementación matemática pura del algoritmo original de atención, este proceso iterativo constituye una pesadilla infraestructural. Para generar cada nueva palabra, el modelo estaría obligado a recalcular forzosamente la proyección completa de las matrices densas de Q, K y V partiendo desde cero para absolutamente todos los tokens de la conversación histórica, desde la primera palabra del prompt hasta el penúltimo token generado. Esta recalculación ingenua induciría una redundancia termodinámica paralizante; la complejidad computacional total del *feed-forward* estallaría exponencialmente con cada iteración.

Por lo que se puede deducir que desde la propuesta original de la arquitectura Transformer (Vaswani et al., 2017), surgió la necesidad crítica de evitar el recálculo de los estados previos

para garantizar la viabilidad comercial de los LLM modernos; ante esto, la ingeniería de software ha estandarizado el uso del *KV Cache* (Caché Clave-Valor). La fundamentación del *KV Cache* reside en la explotación de la propiedad matemática del enmascaramiento causal autorregresivo: las representaciones latentes matriciales de las Claves (K) y Valores (V) calculadas para los tokens pasados son invariantes en el tiempo. Como resultado, la semántica procesada de un token pretérito no se altera por la generación de un token futuro, lo que permite su almacenamiento y reutilización para optimizar la eficiencia computacional (Kwon et al., 2023).

La dinámica operativa se divide en dos fases claramente diferenciadas:

- Fase de Prellenado: Cuando el LLM recibe la solicitud inicial del usuario, el clúster GPU ingiere la totalidad de la secuencia en un único pase altamente paralelizable. Durante esta fase computacionalmente intensiva, la red propaga todos los tokens calculando integralmente sus tensores de K y V. En lugar de descartarlos, siguiendo la propuesta de Vaswani et al. (2017), al finalizar la predicción del primer token de respuesta, la arquitectura transfiere y aloja permanentemente estos tensores dentro de las memorias estáticas VRAM de la GPU, estructurándolos como un bloque de caché histórico continuo a través de todas las capas de atención del modelo.
- **Fase de Decodificación:** Una vez poblado el KV Cache, el modelo genera cada nuevo token de forma iterativa. Para cada paso t, únicamente se calcula el vector de Consulta (Q) del token actual. Este vector se compara contra los tensores K y V almacenados en la VRAM, evitando recalcular el historial completo. Finalmente, el nuevo par (K, V) se anexa al caché para el siguiente paso (Kwon et al., 2023).

El impacto algorítmico del KV Cache es de magnitud fundacional: transforma una operación atencional cuya latencia por iteración poseía una complejidad temporal asintótica cuadrática $\mathcal{O}(n^2)$ en una operación de complejidad estrictamente lineal $\mathcal{O}(n)$ con respecto a la longitud del contexto acumulado (Kwon et al., 2023). Esta herramienta matemática elimina virtualmente el cómputo redundante de operaciones de coma flotante (FLOPs), conteniendo la latencia inter-token en rangos sub-milisegundo y habilitando la telemetría interactiva en tiempo real.

5.3 LIMITACIONES COMPUTACIONALES DE LA ATENCIÓN

5.3.1 LA BARRERA DE LA COMPLEJIDAD CUADRÁTICA ($\mathcal{O}(n^2)$) EN TIEMPO Y MEMORIA

Como se describió en la sección 5.2.1, la Atención de Producto Punto Escalado exige que cada token calcule su alineación con todos los demás tokens de la secuencia, generando una matriz densa de dimensión $L \times L$ (Vaswani et al., 2017). Esta exhaustividad combinatoria impone una complejidad computacional de $\mathcal{O}(L^2 \cdot d_{model})$, lo que representa una barrera crítica para el procesamiento de secuencias largas.

La base matemática de la Atención de Producto Punto Escalado exige que, durante la asimilación del contexto, cada vector individual de la secuencia interactúe algorítmicamente y calcule su alineación semántica cruzada mediante un producto escalar con todos y cada uno de los demás vectores que componen el documento (Vaswani et al., 2017). Esta exhaustividad combinatoria induce la creación forzosa de una matriz densa de interacciones de dimensión $L \times L$, donde L es la cantidad de tokens, es decir, la longitud total de la secuencia del texto.

En términos del análisis asintótico formal, el requerimiento algorítmico de tiempo de procesamiento temporal, evaluado en Operaciones de Punto Flotante o FLOP y de asignación de espacio en memoria durante la fase inicial de pre-entrenamiento o en el prellenado de inferencia escala matemáticamente siguiendo la cota superior $\mathcal{O}(L^2 \cdot d_{model})$ (Vaswani et al., 2017). Para ilustrar las devastadoras implicaciones empíricas de este comportamiento, las operaciones prealimentadas (*MLPs*) de la red y las proyecciones afines escalan de forma gentil y lineal con la longitud de la secuencia. Sin embargo, la operación intrínseca de los cálculos del núcleo de la atención cuadrática se puede aproximar a $\approx 12 \cdot B \cdot L^2 \cdot n_{heads} \cdot d_{heads}$ FLOPs de cómputo termodinámico puro (Korthikanti et al., 2022).

Para secuencias de texto relativamente más cortas, como documentos de 1000 o 2000 tokens, el costo cuadrático de la atención es manejable y queda en segundo plano frente a la enorme carga computacional de las capas perceptrónicas (MLP) del modelo. Sin embargo, el ecosistema corporativo actual demanda capacidades masivas: ingerir repositorios enteros de

código de software, procesar vastos corpus legales y asimilar bibliotecas médicas en un solo prompt. Cuando la longitud de la ventana excede el umbral empírico de aproximadamente $L > 8 \times d_{model}$ (Korthikanti et al., 2022), es decir, cuando el número de tokens de la secuencia supera ocho veces la dimensión del modelo, el costo computacional de la atención comienza a dominar sobre el de las capas perceptrónicas (MLP). Esto frecuentemente equivale a sobrepasar la barrera de los 32K a 64K tokens en modelos de frontera, la curva exponencial dictamina que el algoritmo de atención domina dictatorialmente el presupuesto térmico, los ciclos de reloj y la latencia del clúster. Si un investigador desea expandir un documento hipotético de 100,000 tokens a 200,000 tokens, la carga del prellenado no se duplica, sino que se cuadruplica, llevando a los procesadores al límite térmico, asfixiando los registros de caché L1/L2 SRAM e impidiendo la democratización escalar de estas herramientas sin infraestructuras colosales (Dao et al., 2022).

5.3.2 EL CUELLO DE BOTELLA EN LA MEMORIA DE LA GPU (VRAM) DURANTE LA INFERENCIA

Mientras que el costo temporal $\mathcal{O}(n^2)$ paraliza la asimilación inicial de documentos de gran contexto, la fase persistente de generación está subyugada casi exclusivamente por una métrica infraestructural: el agotamiento y la saturación del ancho de banda de la memoria de video de la Unidad de Procesamiento Gráfico (VRAM). Este fenómeno, conocido en el ámbito del hardware como *Memory-Bandwidth Bound* (Límite por ancho de banda de memoria), es una consecuencia física directa del mecanismo de mitigación del KV Cache, cuya carga computacional y de memoria ha sido caracterizada analíticamente por Pope et al. (2023).

Para comprender la ferocidad cuantitativa de este cuello de botella, el dimensionamiento de las asignaciones de memoria de la VRAM necesarias para albergar el KV Cache obedece analíticamente a la siguiente formulación empírica determinista en bytes (Pope et al., 2023):

$$\text{Consumo Memoria Caché KV} = 2 \times b \times t \times n_{layers} \times n_{heads} \times d_{head} \times p_a.$$

En esta formulación matemática, b representa el tamaño del lote (el número de usuarios concurrentes servidos por el nodo), t denota la sumatoria combinada de la longitud en tokens del prompt inicial más la predicción sintética, n_{layers} designa el número de bloques atencionales

profundos del modelo, n_{heads} es el número de cabezales asociados al par de llaves y valores, d_{head} refleja la dimensionalidad oculta por cabezal y p_a consolida los requerimientos de bytes dictaminados por la precisión numérica. Se utilizan 2 bytes por parámetro en regímenes de precisión media (FP16 o Bfloat16) o 1 byte en entornos cuantizados de 8 bits.

Aplicando esta formulación empírica (Pope et al., 2023) a casos concretos, se revela la cruda realidad del despliegue en la nube. Considerando un modelo fundacional relativamente austero como LLaMA-2-7B (7 mil millones de parámetros), compilado en precisión paramétrica de 16 bits, alojar los tensores estáticos del modelo exige un costo primario de ≈ 14 GB de memoria VRAM pura (Pope et al., 2023). Siguiendo la fórmula expuesta, el alojamiento temporal del caché KV para esta arquitectura dictamina una huella de memoria incremental de aproximadamente 0.5 MB por cada token almacenado en la secuencia de inferencia (Korthikanti et al., 2022; Pope et al., 2023). Si se despliega este mismo modelo en un entorno de producción donde un lote atiende 56 sesiones concurrentes con un contexto asimilado de 512 tokens, la caché requerirá 14 GB adicionales de VRAM (Pope et al., 2023). Al acumular esto a los pesos latentes, el sistema superará invariablemente los 24 GB de VRAM disponibles en una GPU estándar de inferencia (como la NVIDIA A10), colapsando el núcleo en un error crítico de *Out of Memory* (OOM) (Korthikanti et al., 2022). Para arquitecturas titánicas de la magnitud de BLOOM-176B, el costo parasitario dinámico asciende geométricamente a niveles prohibitivos, oscilando en torno a ≈ 4 MB requeridos por token almacenado, haciendo imperativo distribuir los lotes mediante fragmentación (Sharding) paralela de tensores y modelos a lo largo de redes conmutadas de múltiples racks GPU (Korthikanti et al., 2022; Pope et al., 2023).

La crisis infraestructural no radica únicamente en la capacidad estática de la memoria. En cada paso de generación, todo el contenido del KV Cache debe ser transferido desde la memoria principal de la GPU hacia su unidad de cálculo para realizar el producto escalar (Dao et al., 2022). Dado que el ancho de banda de esta transferencia es físicamente limitado, los núcleos de cómputo quedan inactivos esperando los datos, lo que reduce drásticamente la eficiencia y velocidad de generación del modelo (Dao et al., 2022).

5.3.3 DESAFÍOS AL PROCESAR VENTANAS DE CONTEXTO LARGO: DEGRADACIÓN DE RENDIMIENTO Y EL FENÓMENO *LOST IN THE MIDDLE*

La confluencia de infraestructuras distribuidas y parches algorítmicos ingeniosos de compresión, tales como la atención agrupada (*Grouped-Query Attention*) (Ainslie et al., 2023), codificación rotatoria (RoPE) (Su et al., 2024) y paginación en memoria fragmentada (*PagedAttention*) (Kwon et al., 2023), ha permitido a la industria de la IA expandir el límite perimetral de la longitud de las secuencias soportadas. Las arquitecturas vanguardistas contemporáneas aceptan contextualmente desde cientos de miles hasta más de un millón de tokens en un único prompt analítico. No obstante, la viabilidad estructural de introducir un documento gigante en la memoria de un tensor no equivale a que la cognición de la red neuronal pueda asimilarlo, abstraerlo y recuperar su semántica implícita. Numerosos grupos de investigación han expuesto falencias analíticas severas que cuestionan el estado del arte de las capacidades de largo aliento, como el fenómeno de la pérdida de información en los segmentos intermedios del contexto (N. Liu et al., 2024).

Este déficit fue metódicamente examinado y cuantificado por el estudio empírico seminal de N. Liu et al. (2024), acuñando a nivel global el término paradigmático *Lost in the Middle* (Perdido en el Medio). Los investigadores ejecutaron un diseño experimental riguroso sometiendo múltiples familias de LLMs, incluyendo modelos cerrados de frontera y redes de código abierto específicamente entrenadas para ventanas de largo contextoa evaluaciones paramétricas deterministas (N. Liu et al., 2024). Para aislar la capacidad genuina de recuperación de información de la mera memorización del entrenamiento previo, diseñaron flujos de trabajo análogos a la Generación Aumentada por Recuperación (RAG), obligando al modelo a ejecutar tareas mecánicas de Búsqueda de Clave-Valor en largas listas numéricas y Respuestas a Preguntas Multi-Documento (QA) (N. Liu et al., 2024). Al inyectar el párrafo que contenía la respuesta verificable en ubicaciones estocásticamente variables dentro del prompt de contexto extendido, el rendimiento atencional quedó expuesto en su totalidad, revelando una degradación sistemática cuando la información relevante se ubica en el centro del contexto (N. Liu et al., 2024).

Los análisis de N. Liu et al. (2024) revelaron tres patrones de comportamiento consis-

tentes:

- **Sesgo de Primacía (*Primacy Bias*):** El modelo recupera con mayor precisión la información ubicada al inicio del contexto.
- **Sesgo de Recencia (*Recency Bias*):** De forma similar, el modelo retiene mejor la información ubicada al final del contexto, justo antes de generar la respuesta.
- **Decadencia en el Medio (*Lost in the Middle*):** Cuando la información relevante se encuentra en la parte intermedia del contexto, el rendimiento del modelo cae significativamente, llegando incluso a ser inferior al de modelos que no utilizan contexto externo (N. Liu et al., 2024).

Crucialmente, este desmoronamiento cognitivo es observable de forma transversal; el fenómeno persiste incluso en las arquitecturas pre-entrenadas bajo la promesa explícita de soportar mega-contextos extensos (N. Liu et al., 2024). A nivel neuro-matemático, este fallo epistémico no deriva de deficiencias heurísticas transitorias, sino de propiedades latentes y sesgos arraigados a la topología misma de la arquitectura Transformer. Los mecanismos intrínsecos de codificación posicional, aunados a la atención causal autorregresiva y las características distribucionales del lenguaje humano, donde el preámbulo y la conclusión suelen concentrar la mayor densidad informativa, engendran un conflicto probabilístico durante los pases hacia adelante (forward passes) (N. Liu et al., 2024; Su et al., 2024).

Como consecuencia de esta distribución, el hiperplano del Softmax diluye los pesos de importancia matemática de los tokens intermedios a valores casi imperceptibles, priorizando los metadatos iniciales como anclajes de retención y los impulsos finales por su frescura probabilística (N. Liu et al., 2024).

El dictamen empírico derivado del fenómeno «Lost in the Middle» es incontrovertible: la ingeniería perimetral que extiende el alojamiento físico de tokens no se traduce en una fidelidad de recuperación transversal, erosionando el desempeño en despliegues críticos como la auditoría de contratos legales o el tamizado de historiales médicos, donde la asimetría atencional incurre

en errores de alucinación (N. Liu et al., 2024). Este déficit recalca la exigencia metodológica de actualizar los paradigmas de inferencia e implementar protocolos de escrutinio rigurosos sobre toda arquitectura diseñada para contextos largos antes de certificar su aptitud para la asimilación masiva de información (N. Liu et al., 2024).

5.4 ESTRATEGIAS ARQUITECTÓNICAS ALTERNATIVAS AL MECANISMO DE ATENCIÓN ESTÁNDAR

Frente a las limitaciones computacionales descritas en la sección 5.3 (complejidad $\mathcal{O}(n^2)$, saturación del KV-cache y degradación en contextos largos), la literatura del periodo 2020–2025 ha explorado cinco familias arquitectónicas que buscan mitigarlas total o parcialmente: la atención dispersa, la atención lineal, los modelos de espacio de estados (SSM), la mezcla de expertos (MoE) y las arquitecturas híbridas. Cada una opera sobre un eje distinto del problema, ya sea reduciendo el número de pares de tokens evaluados, sustituyendo el softmax por una aproximación de kernel, comprimiendo el historial en un estado recurrente o desacoplando la capacidad paramétrica del costo activo de inferencia. Esta sección expone sus fundamentos conceptuales; el análisis empírico de su adopción y desempeño se desarrolla en el capítulo 7.

5.4.1 ATENCIÓN DISPERSA (SPARSE ATTENTION)

El mecanismo de atención estándar calcula la afinidad entre cada par de tokens de la secuencia, generando una matriz de atención densa de dimensiones $n \times n$. La atención dispersa restringe este cómputo a un subconjunto estructurado de posiciones: ventanas locales, patrones diagonales, tokens globales o combinaciones de estos esquemas (Vaswani et al., 2017). Al reducir el número de pares evaluados de $\mathcal{O}(n^2)$ a $\mathcal{O}(n \cdot k)$, donde $k \ll n$ es el tamaño de la ventana o el número de posiciones seleccionadas, se alivia tanto el costo computacional como el consumo de memoria durante el prefill.

Esta estrategia es especialmente relevante en contextos de inferencia con secuencias largas, donde la saturación del KV-cache constituye el principal cuello de botella. Al limitar las interacciones a patrones locales o de importancia probada, los modelos basados en atención dispersa reducen el ancho de banda requerido en la GPU sin necesidad de alterar la arquitectura

decodificadora base.

5.4.2 ATENCIÓN LINEAL (LINEAR ATTENTION)

La atención lineal reemplaza la operación $\text{softmax}(QK^\top/\sqrt{d_k})V$ por una aproximación basada en funciones de kernel que permite reformular el cálculo como un producto de matrices asociativo: $\phi(Q)(\phi(K)^\top V)$, donde ϕ es un mapa de características positivo (Katharopoulos et al., 2020). Esta reformulación elimina la dependencia cuadrática respecto a la longitud de secuencia n, reduciendo la complejidad temporal y espacial a $\mathcal{O}(n)$ y habilita además una formulación recurrente equivalente que hace la inferencia autoregresiva constante por token.

Su relevancia teórica es elevada: ofrece escalado lineal sin introducir restricciones en el patrón de atención, a diferencia de la atención dispersa. No obstante, la aproximación del softmax tiende a degradar la calidad del modelo en tareas de razonamiento complejo, lo que ha limitado su adopción en LLMs de propósito general.

5.4.3 MODELOS DE ESPACIO DE ESTADOS (SSM)

Los modelos de espacio de estados discretizados, popularizados por la arquitectura Mamba (Gu y Dao, 2023), reformulan el procesamiento de secuencias mediante ecuaciones de la forma $h_t = \bar{A}h_{t-1} + \bar{B}x_t$, $y_t = Ch_t$, donde las matrices $\bar{A}$ y $\bar{B}$ son parámetros aprendidos a partir de una discretización de tiempo continuo. A diferencia de la atención, que mantiene un KV-cache de tamaño $\mathcal{O}(n)$ en VRAM, los SSM comprimen todo el historial en un estado oculto de tamaño fijo, logrando una complejidad de inferencia constante $\mathcal{O}(1)$ por token.

El mecanismo de *selective scan* introducido en Mamba permite que las matrices de transición varíen en función de la entrada, dotando al modelo de selectividad contextual comparable a la atención, pero con un coste de memoria en inferencia radicalmente inferior. Esto los hace especialmente atractivos para despliegue en dispositivos con VRAM restringida y para aplicaciones de contexto extremadamente largo.

5.5 FUNDAMENTOS DE MÉTRICAS DE EVALUACIÓN EN LLMS

5.5.1 MÉTRICAS DE EFICIENCIA COMPUTACIONAL: DEFINICIÓN DE FLOPS, LATENCIA, RENDIMIENTO Y CONSUMO ENERGÉTICO

La evaluación del ecosistema de Modelos de Lenguaje de Gran Escala se divide metodológicamente en dos ramas de validación ortogonales: la robustez del hardware sistémico y la integridad empírica heurística. Debido al masivo gravamen termodinámico y económico subyacente que conlleva el sostenimiento del andamiaje algorítmico, el marco teórico computacional de infraestructura se cuantifica estrictamente a través de un panel de mediciones métricas interrelacionadas altamente precisas: la intensidad computacional absoluta, el desempeño volumétrico generativo de la red, la latencia interactiva y el paradigma de sostenibilidad de carbono emitido (Bari et al., 2025; Luccioni et al., 2023; Pope et al., 2023).

- Carga Operativa Teórica y Práctica (FLOPs y MFU): La unidad de medida microscópica estándar por excelencia en la disciplina de Arquitecturas de Computadoras y Aprendizaje Profundo para diagnosticar y proyectar la huella algorítmica de una red convolucional o atencional es la contabilización total agregada de las Operaciones de Punto Flotante (FLOPs). Cada multiplicación o adición tensorial matricial elemental aporta a esta cuota agregada escalar. De acuerdo con las formulaciones fundacionales de las leyes de escalado empírico y desgloses de hardware exhaustivos por la investigación seminal (Kaplan et al., 2020), las estimaciones heurísticas paramétricas revelan métricas muy contundentes para arquitecturas de Solo Decodificador. En un ciclo iterativo de entrenamiento estándar, que amalgama secuencialmente los recorridos masivos del paso forward, el paso asimétrico backward para el cálculo paramétrico de los gradientes matemáticos y las actualizaciones direccionales de la optimización del modelo, la huella algorítmica aproxima sólidamente $6 \times N \times D$ FLOPs, donde N atestigua el volumen asintótico de parámetros desvinculados de incrustación de la arquitectura neuronal y D designa la longitud total escalar del acervo global de tokens procesado (Kaplan et al., 2020).

  En oposición a esto, durante la vida operativa útil posterior o despliegue de inferencia

estricta, cada iteración atencional forward individual asume teóricamente una merma de solo $2N$ operaciones punto-flotante elementales acumuladas por la adición de cada token subsiguiente, asumiendo que los costos residuales y el overhead puramente atencional de la fase prefill mantengan una cota baja a escalas cortas (Kaplan et al., 2020). Sumado a ello, la comunidad corporativa ha instaurado una meta-métrica sistémica de redimensionamiento, la *Model FLOPs Utilization* (MFU), la cual divide la velocidad teórica termodinámica pico lograda frente al techo operacional físico absoluto declarado por el propio fabricante del silicio del hardware (Chowdhery et al., 2023). Esto funciona como un revelador clínico de si el clúster perimetral se halla padeciendo o siendo asfixiado agudamente por cuellos de botella no procesadores, como latencias críticas de canal inter-red NVLink, saturación extrema de los módulos HBM, entre otros parámetros de paralelización defectuosos, impidiendo desplegar el motor matemático y paralizando la ejecución de matrices en los núcleos (Chowdhery et al., 2023).

- Métricas de Fricción Reactiva Interactiva: Para los subsistemas en los cuales la retroalimentación perimetral directa conversacional inmediata del usuario representa un requisito de servicio excluyente, el Tiempo Hasta el Primer Token (*Time To First Token o TTFT*) gobierna holísticamente la arquitectura. El TTFT abstrae temporalmente la totalidad del período de espera bruto, contabilizado en milisegundos, que media desde la recepción del mensaje contextual de ingreso (prompt) hasta que el modelo emite con éxito la primera fracción léxica o token (Bari et al., 2025).

  Debido a que este delta cronológico aglutina la fase de prefill íntegra de la arquitectura que requiere procesar de golpe el contexto, calcular las proyecciones y saturar la caché vectorial KV en la memoria HBM para la historia completa del documento, además del overhead del enrutador y la espera del planificador, el TTFT es de naturaleza altamente inestable (Pope et al., 2023). Esta métrica crece en correlación cuadrática en respuesta a contextos desproporcionados, asfixiando la sensación humana de inmediatez reactiva si el diseño pre-computacional no ha sido optimizado mediante técnicas de paralelismo o gestión eficiente de memoria (Bari et al., 2025; Pope et al., 2023).

- Métricas Empíricas de Caudal y Rendimiento Continuo: En un espectro contrapuesto, el ecosistema transita a evaluar la estabilidad generativa de estado continuo. La métrica primaria agrupa la relación iterativa cronológica después de superado el prellenado: el Tiempo

por Token de Salida (*Time Per Output Token* o TPOT), frecuentemente utilizado como sinónimo de la Latencia Entre Tokens (*Inter-Token Latency* o ITL) (Samsi et al., 2023). El TPOT refleja directamente el costo algorítmico marginal, dictaminando el intervalo de cálculo puramente escalar para decodificar, anexar el historial en la caché KV y predecir un único token adicional; esto actúa como una representación de la estabilización del núcleo de inferencia sin perturbaciones del prellenado inicial (Pope et al., 2023).

A partir de este estrato, se deducen métricas de flujo como Tokens por Segundo (TPS), las cuales poseen una bifurcación estructural profunda. Por un lado, el User TPS refleja la velocidad del caudal de texto o fluidez experimentada en un socket aislado, calculada matemáticamente como la inversa del TPOT (Samsi et al., 2023). En contraposición, el System TPS (Throughput Global) compila un ratio agregativo de todos los tokens predichos para la pluralidad de usuarios concurrentes en un mismo lote o batch (Pope et al., 2023; Samsi et al., 2023). Esta métrica sirve estrictamente para modelizar el dimensionamiento del servidor en la nube y determinar puntos termodinámicos ciegos de capacidad operativa en infraestructuras a gran escala (Samsi et al., 2023).

- Parámetros Sustentables y Huella de Carbono Ecológica: Con los modelos experimentando saltos masivos de dimensiones parametrales, el gasto energético emerge como una preocupación holística central por los pasivos macroambientales en dióxido de carbono equivalente (CO_2eq) (Luccioni et al., 2023). Para la trazabilidad de la sustentabilidad exaescalar, las lecturas termodinámicas deben transmutarse en la métrica validada de carbono emitido, integrando perfiles de software bajo tres determinantes fundamentales (Luccioni et al., 2023; Patterson et al., 2021):

  1. Potencia Estática e Intensidad Dinámica del Silicio: Audita directamente el volumen de kilovatios consumidos en picos operacionales por los núcleos Tensor, interconexiones HBM y transistores periféricos durante la sesión ininterrumpida (Patterson et al., 2021).
  2. Factor de Eficacia de Uso de la Energía (PUE): El Power Usage Effectiveness es un coeficiente normalizado que dictamina la carga parasitaria agregada (sistemas de enfriamiento e infraestructura del centro de datos) frente al costo estricto del procesamiento del chip (Luccioni et al., 2023). Aplicar un PUE empírico aproxima el cálculo al límite del costo biológico real de la operación.

3. Intensidad de Emisiones Regionales: Actúa modulando la cuantía de gases liberados por la subestación matriz para generar un kilovatio-hora (kWh), lo cual exhibe varianzas astronómicas dependiendo de si la matriz energética local está operada sobre combustibles fósiles o fuentes renovables (Luccioni et al., 2023; Patterson et al., 2021).

Al aplicar estas variables (consumo del hardware multiplicado por el PUE y la intensidad de la zona), la industria modeliza matrices ambientales comparativas ineludibles para certificar la viabilidad ecológica de los despliegues de inteligencia artificial a gran escala (Luccioni et al., 2023).

5.5.2 MÉTRICAS DE CALIDAD Y BENCHMARKS: QUÉ MIDEN LA PERPLEJIDAD Y LOS BENCHMARKS ESTANDARIZADOS

Despojados de las infraestructuras de transistores, la ponderación del constructo de inteligencia latente pre-codificada en los parámetros vectoriales de las redes, así como su heurística de salida, demanda enérgicamente esquemas empíricos transversales taxonómicos. Para ello, es imperativo implementar mediciones y arenas de testeo estandarizadas que certifiquen objetivamente las verdaderas propiedades resolutivas del raciocinio pre-entrenado antes de permitirles escalar responsabilidades en entornos críticos (Burnell et al., 2023; Liang et al., 2022).

- Validación Probabilística Atómica Subyacente: Perplejidad Matemática (*Perplexity*)

  Antes de siquiera aventurar al sistema en la orquestación verbal de lenguaje o la contestación conversacional afín, el motor generativo neuronal requiere una métrica profunda del núcleo basada en el modelaje axiomático del espacio semántico del entrenamiento: la perplejidad (perplexity). Acuñada conceptualmente a partir de los pilares de la teoría general de la información y la abstracción termodinámica (Shannon, 1948), la perplejidad dictamina en términos numéricos la incapacidad probabilística intrínseca y el grado de «sorpresa» o estupor del autómata frente a flujos estructurados de lenguaje natural reservados en arreglos de validación oculta (*hold-out validation corpus*) (Jurafsky y Martin, 2024).

  Desde la perspectiva analítica algebraica, la cota de la perplejidad se calcula y consolida ejecutando la exponenciación de la entropía cruzada (*cross-entropy loss*) promediada

acumulativamente a lo ancho de las derivaciones del ensayo iterativo. Una declinación contundente de este valor certifica fenomenológicamente que la inferencia cognitiva se ha alineado orgánicamente con configuraciones lingüísticas humanas certeras (Jurafsky y Martin, 2024). Pese a ello, esta validación estadística es agnóstica de alucinaciones empíricas y resulta insuficiente de forma autónoma para asegurar la coherencia fáctica final tras el ajuste fino de alineación instruccional, haciendo indispensable recurrir a evaluaciones enciclopédicas estructuradas (Bengio et al., 2003; Jurafsky y Martin, 2024).

- Certificación Enciclopédica del Acervo Epistémico Cognitivo: MMLU

  El estandarte dorado arquitectónico contemporáneo implementado globalmente a manera de evaluación universal del conocimiento ontológico abstracto asimilado por una IA fundacional es el macro-conjunto denominado *Measuring Massive Multitask Language Understanding* (MMLU), conceptualizado por Hendrycks et al. (2021). Desarrollado precisamente para quebrar la ilusión provocada por iteraciones prematuras que inflaban puntuaciones en sistemas como GLUE, MMLU incorporó una inyección exaescalar conformada por 15,908 preguntas complejas de opción múltiple con respuestas de alta jerarquía estocástica no conjeturables sin inducción (Hendrycks et al., 2021).

  Con fines de exposición de deficiencias cruzadas, los cuestionarios diseccionan minuciosamente más de 57 ramas heterogéneas y especialidades académicas, abarcando axiomas en ciencias STEM, directrices normativas en derecho, disyuntivas morales y medicina. Al obligar a la máquina neuronal a actuar sin demostraciones de apoyo (evaluación zero-shot) o auxiliada levemente con micro-inyecciones contextuales (*few-shot*), MMLU sirve para cuantificar irrefutablemente el déficit asimétrico de asimilación del corpus original (Hendrycks et al., 2021). En sus albores, un monolito de 175B de parámetros como GPT-3 original bordeaba una cota ineficiente inferior al $\approx 43{,}9\,\%$ general, exponiendo sesgos gravísimos en la retentiva moral y lógica que posteriormente se abordaron incrementando el ratio dimensional hasta arribar a puntajes de saturación que superan el umbral del $\approx 88{,}5\,\%$ en versiones de vanguardia (Brown et al., 2020; Hendrycks et al., 2021).

- Certificación de Pensamiento Abstracto Matemático Estructurado: GSM8K

  En contraste al tamizado semántico enciclopédico de MMLU, la deducción lógica matemática evidenció deficiencias estructurales en las redes neuronales debido a las propiedades difusas de aproximación de los tokens. Esto estimuló al colectivo científico liderado por

Cobbe et al. (2021) a fraguar la prueba *Grade School Math 8K* (GSM8K). Esta evaluación está orquestada sobre una matriz curada de 8,500 ejercicios matemáticos verbalizados (Word Problems), cuyo contenido simula situaciones cognitivas de nivel pedagógico de escuela secundaria (Cobbe et al., 2021).

La engañosa facilidad aritmética del benchmark descarta redes mediocres de modo inmediato: las especificaciones demandan que la IA interconecte variables extraídas de un párrafo distractor e induzca, mediante el mecanismo de Cadena de Pensamiento (*Chain of Thought* o CoT), el cálculo asimétrico y secuenciado de operaciones algebraicas eslabonadas a través de múltiples pasos de resolución intermedia (Cobbe et al., 2021; Wei et al., 2022). Esta barrera heurística dictaminó la evolución de paradigmas adyacentes, como la obligatoriedad de insertar componentes validadores de puntuación por refuerzo para depurar ramificaciones del árbol de soluciones en vuelo iterativo antes del dictamen final (Cobbe et al., 2021).

- Depuración Axiomática y Heurística Lógica: ARC

  Complementariamente al MMLU y las directrices aritméticas de GSM8K, el Allen Institute for AI instituyó a partir de 2018 la evaluación denominada *AI2 Reasoning Challenge* (ARC), un entorno algorítmico proyectado para diagnosticar los pilares del raciocinio y la conceptualización general (Clark et al., 2018). Su metodología divide la batería evaluativa en un conjunto basal (*Easy Set*) y un segmento crítico perimetral conocido como el Conjunto de Desafío (*Challenge Set*), que contiene más de 7,787 preguntas validadas (Clark et al., 2018).

  El ARC-Challenge anida intgerrogantes donde los baselines históricos, dependientes de esquemas superficiales de correlación y algoritmos de recuperación semántica ingenua, fracasan sistemáticamente. Esto ocurre porque la respuesta vinculada causalmente no reside estadísticamente adyacente a la conformación sintáctica del enunciado (Clark et al., 2018). En este sustrato, es estéril la recuperación por contigüidad; el benchmark determina la capacidad del modelo para forjar relaciones causales dinámicas, aplicar sentido común pragmático y ejecutar lógica multi-paso en escenarios exentos de repetición memorística (Bhakthavatsalam et al., 2021; Clark et al., 2018).

Esta tríada evaluativa, MMLU, GSM8K y ARC, garantiza que el diseño termodinámico

analizado cristalice en simuladores racionales robustos aptos para la interacción global contemporánea.

6. METODOLOGÍA

6.1 PROTOCOLO DE REVISIÓN

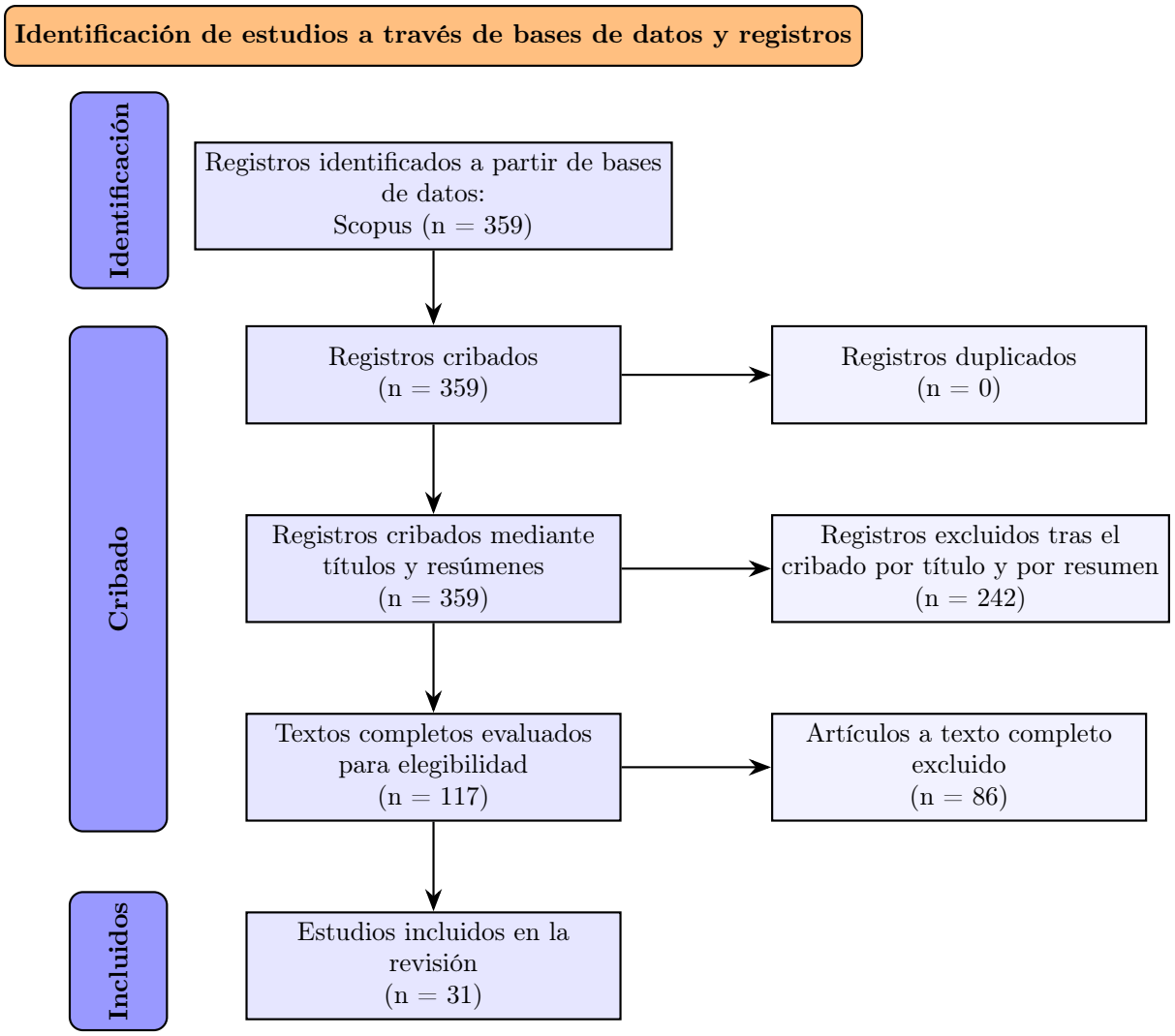
La presente revisión sistemática sigue directamente los lineamientos de la declaración PRISMA 2020 (*Preferred Reporting Items for Systematic reviews and Meta-Analyses*) (**Page2021**), con el fin de garantizar la transparencia, reproducibilidad y trazabilidad del proceso. En primera instancia, se definieron los objetivos del estudio y el alcance metodológico mediante el marco de referencia estructurado en Población, Intervención, Comparación, Resultados y Contexto (PICOC). La población de estudio comprende los estudios empíricos y técnicos sobre modelos de lenguaje de propósito general que cuenten con un billón o más de parámetros. La intervención analizada se centra en los cambios arquitectónicos implementados en los modelos, incluyendo variantes de atención dispersa o lineal, arquitecturas de Mezcla de Expertos y Modelos de Espacio de Estados. Como punto de comparación se establecieron los baselines construidos sobre la arquitectura Transformer clásica de solo decodificador u otras arquitecturas previas evaluadas dentro del mismo trabajo, siempre que las comparaciones se den bajo condiciones declaradas explícitamente. Los resultados a medir se dividen en métricas de eficiencia, tales como operaciones de punto flotante, consumo de memoria, latencia e inferencia y métricas de calidad evaluadas mediante puntuaciones en bancos de pruebas estandarizados como la Perplejidad, MMLU, GSM8K y ARC. Finalmente, el contexto de la revisión se delimita al procesamiento de lenguaje natural generativo en inglés y español, evaluado en entornos académicos e industriales utilizando hardware de procesamiento gráfico disponible públicamente. El protocolo general se resume en la Figura 6.1.

6.2 PREGUNTAS DE INVESTIGACIÓN

Esta revisión sistemática tiene como objetivo central analizar la evolución y la eficiencia de las arquitecturas que plantean alternativas superadoras al mecanismo de atención estándar en los LLMs. Para guiar la extracción y síntesis de datos, se formularon tres preguntas de investigación principales:

Figura 6.1

Protocolo de la revisión sistemática.



Nota. Elaboración propia siguiendo los lineamientos de la declaración PRISMA 2020 (Page et al., 2021).

RQ1: ¿Qué arquitecturas y componentes aparecen con mayor frecuencia en LLMs entre 2020–2025 y cómo evoluciona su proporción por año?

RQ2: ¿Qué rangos y medianas reportan las principales estrategias arquitectónicas en eficiencia y en calidad durante 2020–2025 y cuál es la diferencia mediana entre estrategias comparables bajo condiciones declaradas?

RQ3: ¿Qué proporción de estudios reporta mejoras frente a limitaciones de atención y de qué magnitud por tipo de enfoque?

6.3 DESCRIPCIÓN DE LA ESTRATEGIA DE BÚSQUEDA

La búsqueda principal se realizó en Scopus. Las palabras clave se organizaron en tres bloques temáticos alineados con el marco PICOC:

- **Población (modelos de lenguaje):** "LLM", "large language model", "autoregressive language model", "generative pre-trained transformer"
- **Intervención (arquitecturas):** "architecture", "attention mechanism", "mixture of experts", "recurrent", "sparse attention", "transformer"
- **Resultados (métricas):** "metrics", "ARC", "FLOPs", "GPQA", "MMLU", "VRAM", "benchmark", "computational complexity", "energy consumption"

Estas palabras clave se combinaron mediante la siguiente cadena de búsqueda booleana en el campo `TITLE-ABS-KEY`:

```
TITLE-ABS-KEY ("LLM" OR "large language model" OR "autoregressive language model" OR "generative pre-trained transformer" OR "generative pretrained transformer") AND TITLE-ABS-KEY ("architecture" OR "attention mechanism" OR "mixture of experts" OR "recurrent" OR "sparse attention" OR "transformer") AND TITLE-ABS-KEY ("metrics" OR "ARC" OR "FLOPs" OR "GPQA" OR "MMLU" OR "VRAM" OR "benchmark" OR "computational complexity" OR "energy consumption") AND PUBYEAR >2019 AND PUBYEAR <2026 AND NOT TITLE-ABS ("survey" OR "image" OR "review" OR "clinical") AND ( LIMIT-TO ( SUBJAREA, "COMP" ) OR LIMIT-TO ( SUBJAREA, "ENGI" ) OR LIMIT-TO ( SUBJAREA, "MATH" ) ) AND ( LIMIT-TO ( DOCTYPE, "ar" ) ) AND ( LIMIT-TO ( LANGUAGE, "English" ) OR LIMIT-TO ( LANGUAGE, "Spanish" ) )
```

6.4 SELECCIÓN DE ESTUDIOS

6.5 SELECCIÓN DE ESTUDIOS

El proceso de selección siguió las fases de identificación, cribado e inclusión de la metodología PRISMA. A partir de la búsqueda en Scopus se identificaron 359 registros, sin encontrarse

duplicados. Tras el cribado por título y resumen se excluyeron 242 registros, quedando 117 textos completos evaluados para elegibilidad. De estos, 86 fueron excluidos por no cumplir los criterios establecidos, resultando en un total de **31 estudios incluidos** en la revisión.

Los criterios de inclusión fueron los siguientes:

- Publicación entre 2020 y 2025
- Variantes arquitectónicas explícitas en LLMs
- Condiciones experimentales declaradas claramente
- Modelo con al menos un billón de parámetros

Los criterios de exclusión fueron:

- Arquitecturas no escalables
- Artículos duplicados
- Enfoque únicamente en datos o fine-tuning
- Sin métricas objetivas de eficiencia o calidad
- Idioma distinto al inglés o español
- Modelos cerrados sin métricas auditables
- Modalidades distintas al texto

6.6 SÍNTESIS DE LA LITERATURA Y ANÁLISIS

Para cada uno de los 31 estudios incluidos se completó un formulario de extracción estandarizado registrando: tipo de arquitectura base, año de publicación, métricas de eficiencia reportadas (FLOPs, consumo de memoria, latencia y throughput) y puntuaciones en benchmarks de calidad (Perplejidad, MMLU, GSM8K y ARC).

Con la matriz de registros construida se calcularon las frecuencias de aparición por categoría arquitectónica, identificando que las arquitecturas híbridas dominan el corpus con 18 estudios (58,1 %), seguidas por la atención dispersa y MoE con 6 estudios cada una (19,4 %) y los SSM puros con 1 estudio (3,2 %). La memoria resultó ser la métrica con mayor cobertura (25 estudios), mientras que los benchmarks de calidad presentaron cobertura reducida, con MMLU reportado en apenas 3 estudios. Los resultados completos se presentan organizados según las tres preguntas de investigación en el capítulo siguiente.

7. RESULTADOS Y DISCUSIÓN

Los hallazgos de esta revisión se estructuran en dos secciones principales. Inicialmente, se detalla el perfil del corpus bibliográfico mediante un análisis de su evolución cronológica. Posteriormente, se exponen los resultados técnicos categorizados según las preguntas de investigación establecidas en la metodología.

7.1 PERFIL DE LOS ESTUDIOS INCLUIDOS

Tras la ejecución del protocolo de búsqueda y la aplicación de los criterios de inclusión y exclusión, se consolidó un corpus final de 31 artículos de investigación para el análisis sistémico. El periodo de estudio abarca desde el año 2020 hasta el 2025, evidenciando una tendencia de crecimiento acelerada en la literatura técnica sobre la eficiencia de los modelos de lenguaje de gran escala (LLMs). Mientras que en el periodo inicial (2020-2022) la producción científica no tuvo ningún estudio que cumpla los criterios para pertenecer a este estudio, el volumen de publicaciones tuvo su pico en 2025, concentrando el 83.87 % de los estudios ($n = 26$).

La Tabla 7.1 desagrega esta distribución por año y por categoría arquitectónica. Se observa que el corpus está vacío entre 2020 y 2022, registra una primera incursión aislada en 2023 con un único estudio de atención dispersa y se expande exclusivamente bajo el paradigma híbrido durante 2024 (4 estudios). El año 2025 concentra la totalidad de la diversificación arquitectónica: aparecen los seis estudios de Mezcla de Expertos, el único representante de los modelos de espacio de estados puros y la mayoría de las arquitecturas híbridas (14 de 18). Esta cronología sustenta el análisis temporal desarrollado en la sección 7.4 y evidencia que la validación empírica sistemática de alternativas a la atención estándar es un fenómeno predominantemente de 2025.

7.2 RQ1: ¿QUÉ ARQUITECTURAS Y COMPONENTES APARECEN CON MAYOR FRECUENCIA EN LLMS ENTRE 2020–2025 Y CÓMO EVOLUCIONA SU PROPORCIÓN POR AÑO?

Para dar respuesta a la RQ1, se sistematizaron las arquitecturas y los mecanismos de cómputo reportados en cada uno de los estudios incluidos. Siguiendo la metodología de extrac-

Tabla 7.1

Distribución temporal y arquitectónica del corpus incluido (n = 31)

Año	Sparse Attention	Linear Attention	MoE	SSM	Híbrida	Total	% del corpus
2020	0	0	0	0	0	0	0,0 %
2021	0	0	0	0	0	0	0,0 %
2022	0	0	0	0	0	0	0,0 %
2023	1	0	0	0	0	1	3,2 %
2024	0	0	0	0	4	4	12,9 %
2025	5	0	6	1	14	26	83,9 %
Total	**6**	**0**	**6**	**1**	**18**	**31**	**100 %**

Nota. Elaboración propia a partir de los 31 artículos incluidos en la revisión sistemática.

ción, se identificó que el panorama tecnológico ha migrado desde modelos basados exclusivamente en atención densa hacia estructuras de baja complejidad y alta eficiencia. La Tabla 7.2 resume la matriz de extracción donde se categorizan los estudios según su arquitectura base.

7.2.1 SPARSE ATTENTION (ATENCIÓN DISPERSA)

De los 31 estudios incluidos en esta revisión, seis adoptan la atención dispersa como mecanismo arquitectónico principal: Z. Q. Wang y El Saddik (2023), Kossayakova et al. (2025), Leroux et al. (2025), Z. Liu et al. (2025), Shen y Hu (2025) y Wen et al. (2025). Su distribución temporal es ilustrativa de la trayectoria del enfoque: Z. Q. Wang y El Saddik (2023) constituye el único estudio del corpus publicado en 2023 y las cinco publicaciones restantes se concentran en 2025, lo que sugiere que el enfoque atravesó una fase de validación inicial antes de consolidarse como alternativa práctica frente a la atención densa en escenarios de producción.

El análisis cruzado con las matrices de eficiencia (Tabla 7.3) y de mejoras (Tabla 7.4) revela tres hallazgos consistentes en este subgrupo. En primer lugar, la totalidad de los seis estudios reporta consumo de VRAM y la mediana del grupo (16,0 GB) presenta la mayor reducción absoluta respecto a su baseline entre todas las categorías arquitectónicas del corpus (−2,5 GB mediana), confirmando que la restricción del cómputo a subconjuntos estructurados de tokens se traduce empíricamente en un alivio significativo del cuello de botella del KV-cache. En segundo lugar, cinco de los seis estudios (83 %) reportan latencia y cuatro reportan throughput, una cobertura notablemente superior al promedio del corpus (45 % y 19 % respectivamente), lo que indica que las propuestas de atención dispersa se evalúan predominantemente con métricas de

Tabla 7.2

Matriz de extracción de arquitecturas y componentes identificados ($n = 31$)

Estudio	Sparse Attention	Linear Attention	MoE	SSM	Híbridos
Aly et al. (2025)			✓		
Chen et al. (2024)					✓
Cherednichenko et al. (2025)					✓
Elshaer y Rashed (2025)			✓		
Gao et al. (2024)					✓
Gogineni et al. (2025)					✓
Hu et al. (2025)					✓
Hussain et al. (2025)					✓
Iznaga et al. (2025)				✓	
Jia et al. (2025)			✓		
Khoboko et al. (2025)					✓
Kim y Choi (2025)					✓
Kong et al. (2025)					✓
Kossayakova et al. (2025)	✓				
Kyung et al. (2025)			✓		
Leroux et al. (2025)	✓				
Z. Liu et al. (2025)	✓				
Mahdi et al. (2025)					✓
Martínez et al. (2025)					✓
Meier et al. (2025)			✓		
Ning et al. (2025)					✓
Pau y Aymone (2024)					✓
Pramanik y Bin Husin (2025)					✓
Sheik et al. (2024)					✓
Shen y Hu (2025)	✓				
Udrescu y Popovici (2025)			✓		
Z. Wang et al. (2025)					✓
Z. Q. Wang y El Saddik (2023)	✓				
Wen et al. (2025)	✓				
Wu et al. (2025)					✓
Yu et al. (2025)					✓
Conteo	**6**	**0**	**6**	**1**	**18**

Nota. Elaboración propia a partir de la matriz de extracción RQ1.

inferencia en tiempo real. En tercer lugar, los seis estudios reportan simultáneamente mejoras en eficiencia y en calidad frente al baseline declarado, descartando un trade-off sistemático para este enfoque en las condiciones experimentales revisadas.

La cobertura de benchmarks de calidad estandarizados dentro del subgrupo es, sin embargo, asimétrica: únicamente Leroux et al. (2025) y Wen et al. (2025) reportan perplejidad y solo estos dos estudios incorporan benchmarks de razonamiento (ARC y GSM8K, respectivamente). Esta dispersión replica el patrón general del corpus identificado en la RQ2 y limita la posibilidad de comparar la atención dispersa contra otras estrategias mediante un protocolo unificado de calidad cognitiva. Asimismo, cinco de los seis estudios (todos excepto Z. Q. Wang

y El Saddik (2023)) articulan explícitamente alguna limitación de la atención estándar como motivación central, lo que evidencia que la atención dispersa es, dentro del corpus, una de las líneas de investigación más alineadas discursivamente con la problemática $O(n^2)$.

7.2.2 LINEAR ATTENTION (ATENCIÓN LINEAL)

Ninguno de los 31 estudios incluidos en esta revisión implementa la atención lineal como estrategia arquitectónica principal, lo que constituye en sí mismo un hallazgo significativo: pese a ofrecer la reducción de complejidad más radical entre las alternativas estudiadas ($\mathcal{O}(n^2) \rightarrow \mathcal{O}(n)$ sin restricciones estructurales sobre el patrón de atención), la atención lineal no ha alcanzado validación empírica sistemática en modelos de más de mil millones de parámetros durante el periodo 2020–2025.

Esta ausencia debe interpretarse en contraste con los otros enfoques del corpus: la atención dispersa, que comparte la motivación de aliviar el costo cuadrático pero conserva la estructura del softmax, reúne seis estudios; las arquitecturas híbridas, que también integran bloques de complejidad lineal pero los combinan con capas de atención estándar, suman dieciocho. Esta asimetría sugiere que la comunidad investigadora privilegia las propuestas que preservan la semántica del producto punto escalado, ya sea restringiendo el conjunto de pares evaluados o intercalando bloques alternativos, sobre aquellas que reemplazan íntegramente el softmax por una aproximación de kernel.

Una hipótesis plausible para esta brecha entre potencial teórico y adopción práctica es que los mapas de características empleados para linealizar el softmax aún no preservan la selectividad contextual necesaria para tareas de razonamiento complejo, particularmente bajo los criterios estrictos de comparación cuantitativa con baselines declarados que exige el protocolo de inclusión de esta revisión. Cabe señalar, además, que arquitecturas conceptualmente próximas como los modelos de espacio de estados y diversos esquemas híbridos incorporan elementos de complejidad lineal por vías distintas, lo que indica que la idea subyacente sigue presente en el campo aunque su materialización como atención lineal pura permanezca empíricamente marginal.

7.2.3 STATE SPACE MODELS (SSM)

En el corpus revisado, Iznaga et al. (2025) constituye el único estudio que adopta un SSM puro como arquitectura central (3,2 % del corpus), una representación notablemente exigua dada la visibilidad académica que la familia Mamba (Gu y Dao, 2023) ha alcanzado desde finales de 2023.

El cruce con las matrices de eficiencia (Tabla 7.3) y de mejoras (Tabla 7.4) permite caracterizar el caso de Iznaga et al. (2025): reporta consumo de memoria como única métrica de eficiencia, articula explícitamente las limitaciones de la atención como motivación central y declara mejoras tanto en eficiencia como en calidad frente a su baseline. La ausencia de reporte de latencia, throughput o FLOPs en el único estudio SSM del corpus limita la posibilidad de cuantificar empíricamente la ventaja teórica de complejidad $\mathcal{O}(1)$ por token frente a alternativas comparables.

Esta escasa representación debe leerse en conjunto con la dominancia de las arquitecturas híbridas (18 estudios), varias de las cuales incorporan bloques SSM intercalados con capas de atención estándar bajo el paradigma de Jamba (Lieber et al., 2024). Es decir, el conocimiento técnico de los SSM permea el corpus, pero su integración se materializa predominantemente como componente híbrido más que como sustitución íntegra del Transformer. La fase de consolidación empírica de los SSM puros a escalas superiores a mil millones de parámetros, bajo comparación cuantitativa con baselines declarados, parece aún en curso al cierre del periodo analizado.

7.2.4 MEZCLA DE EXPERTOS (MOE)

La Mezcla de Expertos sustituye las capas feed-forward densas del Transformer por un conjunto de E sub-redes especializadas (expertos) y un enrutador ligero que, para cada token de entrada, selecciona los k expertos más relevantes ($k \ll E$). El cómputo resultante activa únicamente una fracción de los parámetros totales del modelo por cada token procesado, desacoplando la capacidad de representación del costo de inferencia: un modelo MoE puede tener diez veces más parámetros que su contraparte densa sin incrementar proporcionalmente los FLOPs por token (Fedus et al., 2022; Shazeer et al., 2017).

Esta arquitectura es especialmente relevante para escalar la capacidad cognitiva del modelo sin comprometer la latencia interactiva, siempre que la distribución de carga entre expertos sea balanceada. Los desafíos principales residen en el enrutamiento (colapso hacia pocos expertos, inestabilidad del entrenamiento) y en la comunicación inter-GPU en infraestructuras distribuidas.

Seis estudios del corpus aplican MoE como estrategia arquitectónica principal: Aly et al. (2025), Elshaer y Rashed (2025), Jia et al. (2025), Kyung et al. (2025), Meier et al. (2025) y Udrescu y Popovici (2025). Su presencia exclusiva en publicaciones de 2024 y 2025 evidencia que la adopción de MoE en modelos de gran escala se ha intensificado conforme la infraestructura de entrenamiento distribuido se ha democratizado.

7.2.5 ARQUITECTURAS HÍBRIDAS

Las arquitecturas híbridas combinan capas de atención estándar con bloques de naturaleza distinta, típicamente SSMs, convoluciones o capas recurrentes optimizadas, alternando o intercalando ambos tipos dentro del mismo modelo. El paradigma representativo es Jamba (Lieber et al., 2024), que entrelaza capas Transformer y Mamba bajo un esquema MoE, buscando retener la capacidad de razonamiento de la atención global en las capas intercaladas mientras delega el procesamiento secuencial eficiente a los bloques SSM.

La relevancia de este enfoque radica en su pragmatismo: en lugar de reemplazar íntegramente el Transformer, explota la complementariedad entre mecanismos. Las capas de atención preservan la precisión en tareas que requieren recuperación selectiva de información distante, mientras que los bloques alternativos absorben el coste de los tokens intermedios de forma más eficiente. El resultado es un balance configurable entre la exactitud de la atención y la escalabilidad lineal de los modelos alternativos.

Con 18 de los 31 estudios incluidos (58,1 %), las arquitecturas híbridas constituyen la categoría dominante de esta revisión: Chen et al. (2024), Cherednichenko et al. (2025), Gao et al. (2024), Gogineni et al. (2025), Hu et al. (2025), Hussain et al. (2025), Khoboko et al. (2025), Kim y Choi (2025), Kong et al. (2025), Mahdi et al. (2025), Martínez et al. (2025), Ning et al.

(2025), Pau y Aymone (2024), Pramanik y Bin Husin (2025), Sheik et al. (2024), Z. Wang et al. (2025), Wu et al. (2025) y Yu et al. (2025). Su predominancia refleja una tendencia de la comunidad investigadora a construir sobre la infraestructura Transformer existente en lugar de reemplazarla, adoptando innovaciones de forma incremental y validada.

7.3 RQ2: ¿QUÉ RANGOS Y MEDIANAS REPORTAN LAS PRINCIPALES ESTRATEGIAS ARQUITECTÓNICAS EN EFICIENCIA Y EN CALIDAD DURANTE 2020–2025 Y CUÁL ES LA DIFERENCIA MEDIANA ENTRE ESTRATEGIAS COMPARABLES BAJO CONDICIONES DECLARADAS?

Para responder a la RQ2, se extrajeron los valores numéricos de ocho métricas de eficiencia y calidad reportadas en los 31 estudios incluidos: FLOPs, consumo de memoria, latencia, throughput, perplejidad, MMLU, GSM8K y ARC. La Tabla 7.3 sintetiza la cobertura por estudio, evidenciando una marcada heterogeneidad en las condiciones de reporte: la memoria es la métrica más documentada ($n = 25$; 80,6 % del corpus), seguida por la latencia ($n = 14$; 45,2 %), mientras que los benchmarks de calidad estandarizados presentan una cobertura reducida, con MMLU reportado en apenas tres estudios y perplejidad, GSM8K y ARC en dos cada uno. Esta dispersión refleja la ausencia de un protocolo de evaluación unificado en la literatura: cada trabajo selecciona las métricas más favorables para su propuesta, lo que limita la comparabilidad directa entre estrategias arquitectónicas y obliga a interpretar los rangos y medianas resultantes con cautela metodológica.

Tabla 7.3

Matriz de extracción de métricas reportadas por estudio (RQ2, $n = 31$)

Estudio	FLOPs	Memoria	Latencia	Throughput	Perplexity	MMLU	GSM8K	ARC
Aly et al. (2025)								
Chen et al. (2024)		✓	✓	✓			✓	
Cherednichenko et al. (2025)								
Elshaer y Rashed (2025)	✓	✓	✓					
Gao et al. (2024)	✓	✓						
Gogineni et al. (2025)		✓						
Hu et al. (2025)	✓	✓	✓					
Hussain et al. (2025)		✓	✓	✓				
Iznaga et al. (2025)		✓						
Jia et al. (2025)		✓						
Khoboko et al. (2025)		✓						
Kim y Choi (2025)		✓	✓					
Kong et al. (2025)	✓							
Kossayakova et al. (2025)		✓						
Kyung et al. (2025)		✓						
Leroux et al. (2025)		✓	✓	✓	✓			✓
Z. Liu et al. (2025)	✓	✓	✓	✓				
Mahdi et al. (2025)		✓	✓					
Martínez et al. (2025)			✓	✓				
Meier et al. (2025)	✓	✓	✓					
Ning et al. (2025)		✓	✓					
Pau y Aymone (2024)								
Pramanik y Bin Husin (2025)		✓						
Sheik et al. (2024)		✓	✓					
Shen y Hu (2025)		✓	✓					
Udrescu y Popovici (2025)		✓				✓		
Z. Q. Wang y El Saddik (2023)		✓						
Z. Wang et al. (2025)						✓		✓
Wen et al. (2025)		✓	✓	✓	✓		✓	
Wu et al. (2025)		✓				✓		
Yu et al. (2025)		✓						
Conteo	**6**	**25**	**14**	**6**	**2**	**3**	**2**	**2**

Nota. Elaboración propia a partir de la matriz de extracción RQ2.

7.3.1 LATENCIA E INFERENCIA (THROUGHPUT)

Las definiciones operativas del Tiempo Hasta el Primer Token (TTFT) y del Tiempo por Token de Salida (TPOT), junto con sus derivados de throughput agregado, fueron establecidas

en la sección 5.5.1 del marco teórico. Esta subsección analiza su patrón de reporte y los hallazgos empíricos extraídos del corpus.

De los 31 estudios revisados, 14 (45,2 %) reportan alguna medida de latencia y 6 (19,4 %) reportan throughput de generación. La distribución por categoría arquitectónica revela una asimetría informativa: la atención dispersa exhibe la mayor tasa relativa de reporte de latencia (4 de 6 estudios, 66,7 %) y de throughput (3 de 6, 50 %), seguida por las arquitecturas híbridas (8 de 18 en latencia, 44 %; 3 de 18 en throughput, 16,7 %), MoE (2 de 6 en latencia, 33 %) y el único estudio SSM, que no reporta ninguna métrica de inferencia temporal. Este patrón es consistente con el perfil de motivación discursiva de la atención dispersa, que articula explícitamente la latencia del prefill como problema central.

Entre los estudios que documentan mejoras cuantitativas en tokens por segundo al apartarse de la atención densa se encuentran Z. Liu et al. (2025), Chen et al. (2024), Martínez et al. (2025), Hussain et al. (2025) y Leroux et al. (2025). Wen et al. (2025) reporta el throughput agregado de mayor magnitud del corpus, aunque bajo condiciones de procesamiento en lote (*batch*) que no son directamente comparables con la generación interactiva de un único usuario, donde el TPOT individual gobierna la experiencia percibida.

La heterogeneidad de hardware empleado para las mediciones (desde GPUs de consumo hasta clústeres H100) y la ausencia de un protocolo unificado de reporte (longitud de prompt, tamaño de batch, precisión numérica) constituyen el principal obstáculo metodológico para extraer una mediana de latencia comparativa entre estrategias. Esta limitación replica el patrón observado a lo largo de toda la RQ2: cada estudio selecciona las condiciones experimentales más favorables a su propuesta, lo que dificulta el meta-análisis sistemático del balance latencia–throughput entre arquitecturas alternativas.

7.3.2 USO DE MEMORIA Y VRAM

El consumo de memoria de video durante la inferencia en LLMs está determinado por tres componentes principales: los pesos del modelo, las activaciones intermedias y el KV-cache. Este último almacena proyecciones de clave y valor ya calculadas para todo el historial de la

conversación, evitando su recómputo en cada paso autoregresivo (Pope et al., 2023). Su tamaño crece linealmente con la longitud de la secuencia y el número de capas, convirtiéndose en el principal factor de saturación de VRAM durante sesiones de contexto largo.

La reducción de la huella de memoria tiene consecuencias directas sobre la accesibilidad de los modelos: permite su despliegue en hardware de consumo con capacidades de VRAM restringidas (8–24 GB), habilita el procesamiento de ventanas de contexto más amplias sin partición de secuencia y reduce el coste operativo de los servicios en la nube al disminuir la clase de GPU requerida. Los SSM, en particular, eliminan el KV-cache al comprimir el historial en un estado recurrente de tamaño fijo, logrando consumo de memoria constante $O(1)$ por token durante la inferencia (Gu y Dao, 2023).

La memoria es la métrica con mayor cobertura en el corpus: 25 de los 31 estudios la reportan. Los modelos híbridos presentan la mediana más baja del conjunto (8,0 GB), seguidos por los de atención dispersa (16,0 GB) y MoE (24,6 GB). En términos de diferencia respecto al baseline, las arquitecturas de atención dispersa exhiben la mayor reducción mediana ($-2{,}5$ GB), seguidas por los híbridos ($-0{,}75$ GB), mientras que MoE no registra reducción media apreciable. Los estudios que analizan explícitamente el contraste de VRAM entre Transformers estándar y sus alternativas incluyen Shen y Hu (2025), Kim y Choi (2025), Leroux et al. (2025), Iznaga et al. (2025) y Gogineni et al. (2025).

7.3.3 COMPLEJIDAD COMPUTACIONAL (FLOPS)

Las operaciones de punto flotante (FLOPs) constituyen la medida canónica del coste computacional de un modelo, tanto durante el entrenamiento ($\approx 6ND$ FLOPs, donde N es el número de parámetros y D el volumen de tokens) como en la inferencia ($\approx 2N$ FLOPs por token generado) (Kaplan et al., 2020). En la atención estándar, la multiplicación $QK^\top$ exige $O(n^2 d)$ operaciones por capa, donde n es la longitud de la secuencia y d la dimensión de las cabezas. Esta dependencia cuadrática hace que, al duplicar la longitud del contexto, el coste de atención se cuadruplique, limitando severamente el escalado hacia ventanas de contexto masivas.

Las arquitecturas alternativas rompen esta barrera de formas distintas: la atención dis-

persa reduce los pares evaluados a $O(nk)$ con $k \ll n$; los SSM reformulan el cómputo como una convolución o recurrencia con coste $O(n)$; y MoE reduce los FLOPs activos por token al activar únicamente una fracción de los expertos. La comparación empírica de FLOPs entre propuesta y baseline permite cuantificar el ahorro computacional real bajo condiciones controladas.

Solo 6 de los 31 estudios reportan FLOPs de forma explícita y comparable: Z. Liu et al. (2025), Kong et al. (2025), Gao et al. (2024), Elshaer y Rashed (2025), Hu et al. (2025) y Meier et al. (2025). La baja cobertura de esta métrica en el corpus refleja la heterogeneidad de las condiciones experimentales reportadas: diferencias en el tamaño del modelo, longitud de secuencia de evaluación y unidades empleadas (GFLOPs vs. TFLOPs) dificultan la comparación directa entre estudios.

7.3.4 CALIDAD PREDICTIVA Y BENCHMARKS DE RENDIMIENTO

La evaluación de la calidad cognitiva de un LLM requiere métricas que trasciendan la fluidez superficial y verifiquen capacidades de razonamiento generalizables. La perplejidad mide la incertidumbre media del modelo sobre secuencias de validación no vistas: valores menores indican mejor modelado del lenguaje, aunque son agnósticos respecto a la coherencia factual (Jurafsky y Martin, 2024). Los benchmarks MMLU (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021) y ARC (Clark et al., 2018) evalúan, respectivamente, conocimiento enciclopédico multidisciplinar, razonamiento matemático secuencial y abstracción lógica causal, proporcionando una cobertura complementaria de las capacidades del modelo.

Demostrar que las ganancias de eficiencia no degradan la calidad cognitiva es un requisito crítico para la adopción de arquitecturas alternativas: un modelo más rápido o más frugal en memoria carece de valor práctico si sus capacidades de razonamiento se deterioran apreciablemente frente al Transformer denso de referencia. Esta tensión entre eficiencia y calidad constituye el trade-off central que la literatura del periodo 2020–2025 busca resolver.

La cobertura de benchmarks de calidad en el corpus es, sin embargo, reducida. La perplejidad es reportada por Wen et al. (2025) y Leroux et al. (2025). El MMLU aparece en Udrescu y Popovici (2025) (78,0 %), Wu et al. (2025) (69,4 %) y Z. Wang et al. (2025) (41,8 %). GSM8K

es evaluado por Chen et al. (2024) (62,2 %) y Wen et al. (2025) (52,1 %) y ARC por Z. Wang et al. (2025) (44,5 %) y Leroux et al. (2025) (43,9 %). La escasa representación de benchmarks estandarizados en el corpus sugiere que una fracción significativa de los estudios valida sus propuestas mediante métricas propias o específicas de tarea, lo que dificulta la comparación sistemática entre enfoques arquitectónicos.

7.3.5 EFICIENCIA ENERGÉTICA Y HUELLA DE CARBONO

El consumo eléctrico de los LLMs durante el entrenamiento y la inferencia se estima integrando la potencia activa del hardware (núcleos Tensor, interconexiones HBM), el coeficiente de eficiencia del centro de datos (PUE) y la intensidad de emisiones de la red eléctrica local, expresando el resultado en kilogramos de CO_2 equivalente (CO_2eq) (Luccioni et al., 2023; Patterson et al., 2021). Modelos entrenados en regiones con matrices energéticas de alta participación renovable pueden presentar huellas de carbono hasta dos órdenes de magnitud menores que sus equivalentes operados sobre combustibles fósiles, aun con idéntica arquitectura.

La sostenibilidad ambiental ha emergido como un criterio de evaluación de primer orden en el diseño de LLMs, particularmente ante el crecimiento exponencial del coste energético asociado al escalado de parámetros y datos (Patterson et al., 2021). Las arquitecturas que reducen los FLOPs activos por token o la memoria requerida no solo disminuyen el coste operativo económico, sino también la huella térmica y de carbono del sistema, lo que resulta especialmente relevante en despliegues a gran escala o en dispositivos de consumo con autonomía de batería limitada.

En el corpus revisado, cuatro estudios incorporan el consumo energético como factor explícito de evaluación: Z. Liu et al. (2025), Gogineni et al. (2025), Leroux et al. (2025) y Martínez et al. (2025). Su presencia minoritaria (13 % del corpus) indica que la cuantificación de la huella de carbono no se ha consolidado aún como práctica estándar en la publicación de nuevas arquitecturas de LLMs, a pesar de su creciente relevancia metodológica y regulatoria.

7.4 RQ3: ¿QUÉ PROPORCIÓN DE ESTUDIOS REPORTA MEJORAS FRENTE A LIMITACIONES DE ATENCIÓN Y DE QUÉ MAGNITUD POR TIPO DE ENFOQUE?

Para responder a la RQ3, se analizó cada uno de los 31 estudios incluidos en tres dimensiones: si el trabajo motiva explícitamente su propuesta en alguna limitación del mecanismo de atención estándar (complejidad cuadrática, saturación del KV-cache, degradación en contextos largos, latencia o consumo energético), si reporta mejora cuantitativa frente a su baseline en eficiencia y si lo hace en calidad cognitiva. La Tabla 7.4 sintetiza el resultado por estudio en estas tres dimensiones.

Los resultados revelan una asimetría notable. Por un lado, 18 de los 31 estudios (58,1 %) articulan explícitamente una limitación de la atención como motivación central. Por otro, 29 (93,5 %) reportan mejoras medibles en al menos una métrica de eficiencia y 28 (90,3 %) en al menos una métrica de calidad.

Esta brecha entre quienes declaran motivación en la atención y quienes efectivamente reportan mejoras sugiere que una fracción del corpus aborda optimizaciones de eficiencia sin enmarcarlas explícitamente en la problemática $\mathcal{O}(n^2)$. Esto ocurre bien porque su contribución es de naturaleza distinta, bien porque la limitación se asume implícitamente como contexto compartido del campo.

Tabla 7.4

Proporción de estudios que abordan limitaciones de atención y reportan mejoras (RQ3, $n = 31$)

Estudio	Aborda limitaciones	Mejora eficiencia	Mejora calidad
Aly et al. (2025)	✓	✓	✓
Chen et al. (2024)	✓	✓	
Cherednichenko et al. (2025)	✓	✓	✓
Elshaer y Rashed (2025)		✓	✓
Gao et al. (2024)	✓	✓	✓
Gogineni et al. (2025)	✓	✓	✓
Hu et al. (2025)	✓	✓	✓
Hussain et al. (2025)		✓	✓
Iznaga et al. (2025)	✓	✓	✓
Jia et al. (2025)		✓	✓
Khoboko et al. (2025)			✓
Kim y Choi (2025)	✓	✓	✓
Kong et al. (2025)		✓	✓
Kossayakova et al. (2025)	✓	✓	✓
Kyung et al. (2025)			
Leroux et al. (2025)	✓	✓	✓
Z. Liu et al. (2025)	✓	✓	✓
Mahdi et al. (2025)		✓	✓
Martínez et al. (2025)	✓	✓	✓
Meier et al. (2025)		✓	✓
Ning et al. (2025)		✓	✓
Pau y Aymone (2024)	✓	✓	
Pramanik y Bin Husin (2025)	✓	✓	✓
Sheik et al. (2024)		✓	✓
Shen y Hu (2025)	✓	✓	✓
Udrescu y Popovici (2025)		✓	✓
Z. Q. Wang y El Saddik (2023)		✓	✓
Z. Wang et al. (2025)	✓	✓	✓
Wen et al. (2025)	✓	✓	✓
Wu et al. (2025)		✓	✓
Yu et al. (2025)	✓	✓	✓
Conteo	**18**	**29**	**28**

Nota. Elaboración propia a partir de la matriz de extracción RQ3.

7.4.1 LA ERA DE LA CONSOLIDACIÓN (2020–2022): HEGEMONÍA DE LA ATENCIÓN CUADRÁTICA

Entre 2020 y 2022, el ecosistema de investigación en LLMs estuvo marcado por el escalamiento por fuerza bruta sobre la arquitectura Transformer estándar: más parámetros, más datos y más cómputo, siguiendo las leyes de escalado formuladas por Kaplan et al. (2020) y ejecutadas masivamente en modelos como GPT-3 (Brown et al., 2020). La atención de producto punto escalado, con su complejidad $O(n^2)$, era el mecanismo dominante e indiscutido y la respuesta al cuello de botella de memoria consistía principalmente en aumentar la capacidad de las GPUs o fragmentar las secuencias. Técnicas como Flash Attention (Dao et al., 2022) comenzaron a optimizar el acceso a memoria sin alterar la semántica del cálculo, representando una mejora incremental dentro del paradigma existente antes que una ruptura arquitectónica.

Ninguno de los 31 estudios incluidos en esta revisión fue publicado en este periodo. Esta ausencia no es accidental: los criterios de inclusión del protocolo exigen la comparación explícita con un baseline de atención estándar bajo condiciones declaradas, un rigor metodológico que la literatura de 2020–2022 aún no estandarizaba. El periodo funciona, por tanto, como la línea de base conceptual sobre la que se mide el progreso documentado en los años siguientes.

7.4.2 EL PUNTO DE INFLEXIÓN (2023): LA TRANSICIÓN SUB-CUADRÁTICA

El año 2023 marca la primera señal de transición en el corpus: un único estudio satisface los criterios de inclusión, Z. Q. Wang y El Saddik (2023), que implementa atención dispersa como mecanismo central para reducir la complejidad cuadrática en la fase de prefill. Su presencia aislada en el corpus es coherente con el contexto del campo: 2023 es el año en que se publican los trabajos que definirán el siguiente ciclo, entre ellos Mamba (Gu y Dao, 2023) en diciembre, que introduce el mecanismo de selective scan como alternativa SSM al Transformer y las primeras versiones de modelos MoE a escala de producción.

La urgencia que motiva estos desarrollos es el crecimiento de las ventanas de contexto exigidas por aplicaciones reales, desde el análisis de contratos legales hasta la ingesta de repositorios de código completos, frente a las cuales la saturación cuadrática del KV-cache constituía

un obstáculo operativo insalvable con el hardware disponible. El corpus refleja este momento como una bisagra: el año en que la literatura académica comienza a plantear rupturas, aunque su validación empírica a escala aún no ha alcanzado la densidad suficiente para generar múltiples estudios que cumplan los criterios de inclusión.

7.4.3 EL CAMBIO DE PARADIGMA (2024): CONSOLIDACIÓN EXPERIMENTAL DE LAS ARQUITECTURAS HÍBRIDAS

En 2024, el corpus registra cuatro estudios, todos clasificados como arquitecturas híbridas: Chen et al. (2024), Gao et al. (2024), Pau y Aymone (2024) y Sheik et al. (2024). Esta concentración exclusiva en el paradigma híbrido refleja la respuesta pragmática de la comunidad investigadora a los avances teóricos de 2023: antes de reemplazar íntegramente la atención estándar, los investigadores optan por intercalar bloques alternativos, SSMs o capas recurrentes optimizadas, dentro de la arquitectura Transformer existente, validando empíricamente su comportamiento frente a baselines bien establecidos.

Es relevante señalar que las arquitecturas puramente SSM y los modelos MoE a gran escala, cuya irrupción teórica ocurrió en 2023–2024 con trabajos como Mamba (Gu y Dao, 2023) y Mixtral (Jiang et al., 2024), no alcanzan en este corpus representación con fecha de publicación de 2024. Esto sugiere que su validación empírica bajo los estrictos criterios de inclusión del protocolo, comparación cuantitativa con baseline en condiciones declaradas en modelos de más de mil millones de parámetros, requirió un ciclo adicional de maduración antes de generar literatura publicable que los documentara de forma sistemática.

7.4.4 TENDENCIAS CONSOLIDADAS Y PERSPECTIVAS (2025)

El año 2025 concentra 26 de los 31 estudios del corpus (83,9 %), confirmando el pico de actividad investigadora predicho por la curva de adopción tecnológica. La distribución arquitectónica de este subconjunto revela una convergencia hacia tres líneas de desarrollo simultáneas: las arquitecturas híbridas dominan con 14 estudios (Cherednichenko et al. (2025), Gogineni et al. (2025), Hu et al. (2025), Hussain et al. (2025), Khoboko et al. (2025), Kim y Choi (2025), Kong et al. (2025), Mahdi et al. (2025), Martínez et al. (2025), Ning et al. (2025), Pramanik

y Bin Husin (2025), Z. Wang et al. (2025), Wu et al. (2025), Yu et al. (2025)); la Mezcla de Expertos consolida su presencia con 6 estudios (Aly et al. (2025), Elshaer y Rashed (2025), Jia et al. (2025), Kyung et al. (2025), Meier et al. (2025), Udrescu y Popovici (2025)); y la atención dispersa mantiene vigencia con 5 publicaciones (Z. Liu et al. (2025), Kossayakova et al. (2025), Leroux et al. (2025), Shen y Hu (2025), Wen et al. (2025)). Los SSM puros aparecen únicamente en Iznaga et al. (2025).

Dos tendencias transversales caracterizan esta cohorte. La primera es el diseño orientado al hardware (*hardware-aware*): múltiples estudios optimizan sus arquitecturas teniendo en cuenta las jerarquías de memoria específicas de las GPUs modernas, reduciendo el ancho de banda requerido entre HBM y SRAM. La segunda es la incorporación progresiva de la eficiencia energética como criterio de evaluación explícito junto a las métricas de latencia y calidad, un indicador de que la sostenibilidad del despliegue a escala ha entrado en la agenda de la comunidad investigadora como restricción de diseño y no solo como consideración posterior.

La Figura 7.1 sintetiza la evolución temporal del corpus por categoría arquitectónica, evidenciando la concentración del 83,9 % de los estudios en 2025 y la diversificación simultánea de enfoques que caracteriza ese año.

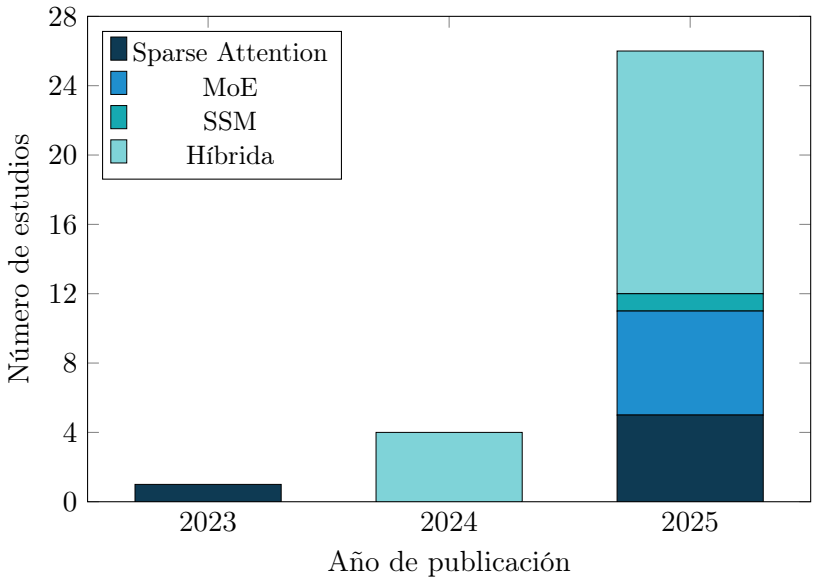


Figura 7.1

Evolución temporal del corpus por categoría arquitectónica (2023–2025, n = 31)

Nota. Elaboración propia a partir de la matriz de extracción RQ1. No se incluyen los años 2020–2022 por no registrar estudios que cumplieran los criterios de inclusión, ni la categoría Linear Attention por no contar con representación en el corpus.

8. CONCLUSIONES

La presente revisión sistemática, estructurada bajo la declaración PRISMA 2020 y circunscrita al periodo 2020–2025, consolidó un corpus de 31 artículos empíricos sobre arquitecturas de LLMs que implementan alternativas al mecanismo de atención estándar. Sus hallazgos responden de forma directa a los tres objetivos planteados.

Respecto al primer objetivo, la estrategia dominante del periodo consiste en hibridar el Transformer en lugar de sustituirlo: 18 de los 31 estudios (58,1 %) corresponden a arquitecturas híbridas, frente al 19,4 % de atención dispersa y de Mezcla de Expertos (6 estudios cada una) y un único estudio de SSM puro (3,2 %). El hallazgo más revelador es una ausencia: ningún estudio implementa atención lineal como estrategia principal, lo que indica que sus aproximaciones de kernel aún no preservan la calidad de razonamiento exigida a escalas superiores a mil millones de parámetros.

Respecto al segundo objetivo, la memoria es la única métrica con cobertura suficiente (25/31 estudios) para una comparación interarquitectónica robusta, mientras que los benchmarks de calidad (MMLU, GSM8K, ARC) aparecen en un máximo de tres estudios cada uno. Esta subdocumentación de la calidad cognitiva constituye un déficit estructural del campo que impide afirmar cuantitativamente si la migración hacia arquitecturas alternativas preserva o degrada la capacidad de razonamiento, y revela la urgencia de un protocolo de evaluación común con métricas mínimas obligatorias bajo condiciones declaradas.

Respecto al tercer objetivo, las arquitecturas más allá de la atención estándar mejoran de forma mayoritaria tanto la eficiencia (29/31; 93,5 %) como la calidad (28/31; 90,3 %) frente a sus baselines, lo que descarta un trade-off sistemático e ineludible entre ambas dimensiones dentro de las condiciones documentadas. La asimetría frente al 58,1 % que declara explícitamente una motivación en las limitaciones de la atención sugiere que la problemática $O(n^2)$ se asume ya como contexto compartido del campo.

Estos resultados deben interpretarse considerando las limitaciones del estudio: la bús-

queda se circunscribió a Scopus, se restringió a modelos de más de mil millones de parámetros y la heterogeneidad de las condiciones experimentales limita la comparabilidad directa de las medianas. De ellas se derivan tres direcciones futuras prioritarias: estandarizar un protocolo de reporte de métricas comparables, validar empíricamente los SSM a escalas de 7B parámetros y superiores, e incorporar la eficiencia energética como métrica de reporte obligatorio.

En síntesis, el estado del arte de las arquitecturas de LLMs entre 2020 y 2025 describe una transición pragmática: el Transformer permanece como base, profundamente modificado mediante la incorporación incremental de bloques alternativos. La evidencia confirma que es posible superarlo en condiciones controladas; la pregunta que define el siguiente ciclo de investigación es bajo qué condiciones esa superioridad se sostiene a escala de producción, con hardware heterogéneo y en tareas de razonamiento de alta complejidad.

9. REFERENCIAS

AI@Meta, Dubaih, A., Jauhri, A., Pandey, A. et al. (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*. https://arxiv.org/pdf/2407.21783

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Real, F. y Dhariwal, J. (2023). GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. *arXiv preprint arXiv:2305.21958*. https://arxiv.org/pdf/2305.13245

Almeida, M. (2020). *Aprendizaje en el área de matemáticas: una propuesta pedagógica desde la gamificación* (inf. téc.) (Tesis de posgrado). Pontificia Universidad Católica del Ecuador, Quito.

Aly, W. M., Soliman, T. H. A. y AbdelAziz, A. M. (2025). Cross-Domain Evaluation of Large Language Models for Abstractive Text Summarization: An Empirical Perspective. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *16*(6). https://doi.org/10.14569/IJACSA.2025.0160695

Arias, F. (2012). *El proyecto de investigación. Introducción a la metodología científica*. Editorial Episteme.

Bari, A., Hegde, P. y de Veciana, G. (2025). Optimal Scheduling Algorithms for LLM Inference: Theory and Practice. *arXiv preprint arXiv:2508.01002*. https://arxiv.org/pdf/2508.01002

Bartle, R. e Islas, D. (2004). *Introducción al Análisis Matemático de una variable* (2.ª ed.). Limusa Wiley.

Bartle, R. y Sherbet, D. (2004). *Introducción al Análisis Matemático de una variable* (2.ª ed.). Limusa Wiley.

Bengio, Y., Ducharme, R., Vincent, P. y Janvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, *3*, 1137-1155. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

Bengio, Y., LeCun, Y. e Hinton, G. (2021). Deep Learning for AI. *Communications of the ACM*, *64*(7), 58-65. https://doi.org/10.1145/3448250

Bhakthavatsalam, S., Khashabi, D., Khot, T., Dalvi, B., Richardson, K., Sabharwal, A., Schoenick, C., Tafjord, O. y Clark, P. (2021). Think you have Solved Direct-Answer Question

Answering? Try ARC-DA, the Direct-Answer AI2 Reasoning Challenge. *arXiv preprint arXiv:2102.03315*. https://arxiv.org/pdf/2102.03315

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E. et al. (2021). *On the Opportunities and Risks of Foundation Models* (inf. téc.). Stanford Center for Research on Foundation Models (CRFM). https://crfm.stanford.edu/assets/report.pdf

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language Models are Few-Shot Learners.

Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., Sohl-Dickstein, J., Mitchell, M., Kiela, D., Shanahan, M. et al. (2023). Rethink reporting of evaluation results in AI. *Science*, *380*(6641), 136-138. https://doi.org/10.1126/science.adf6369

Chen, H., Zhang, J., Du, Y., Xiang, S., Yue, Z., Zhang, N., Cai, Y. y Zhang, Z. (2024). Understanding the Potential of FPGA-based Spatial Acceleration for Large Language Model Inference [Cited by: 55; All Open Access; Bronze Open Access]. *ACM Transactions on Reconfigurable Technology and Systems*, *18*(1), 1-29. https://doi.org/10.1145/3656177

Cherednichenko, O., Herbert, A. y Poptsova, M. (2025). Benchmarking DNA large language models on quadruplexes [Cited by: 0; All Open Access; Gold Open Access; Green Open Access]. *Computational and Structural Biotechnology Journal*, *27*, 992-1000. https://doi.org/10.1016/j.csbj.2025.03.007

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S. et al. (2023). PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, *24*(240), 1-113. https://www.jmlr.org/papers/volume24/22-1144/22-1144.pdf

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C. y Tafjord, O. (2018). Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*. https://arxiv.org/pdf/1803.05457

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C. y Schulman, J. (2021). Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*. https://arxiv.org/pdf/2110.14168

Coronel, M. (2020). *Guía didáctica para la enseñanza de Matemática de los estudiantes de octavo año de la escuela de educación básica superior para personas con escolaridad inconclusa*

Tarqui, modalidad semipresencial intensiva. (inf. téc.) (Tesis de posgrado). Universidad Central del Ecuador, Quito.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A. y Ré, C. (2022). FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. *arXiv preprint arXiv:2205.14135*. https://arxiv.org/pdf/2205.14135

Devlin, J., Chang, M.-W., Lee, K. y Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*. https://arxiv.org/pdf/1810.04805

Elshaer, Z. y Rashed, E. A. (2025). CURE: Confidence-Driven Unified Reasoning Ensemble Framework for Medical Question Answering [Cited by: 0; All Open Access; Gold Open Access; Green Open Access]. *Big Data and Cognitive Computing*, *9*(12). https://doi.org/10.3390/bdcc9120299

Fedus, W., Zoph, B. y Shazeer, N. (2022). Switch Transformers: Scaling to Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, *23*(120), 1-39. https://arxiv.org/abs/2101.03961

Gaitán, V. (2013). *Gamificación: el aprendizaje divertido.* https://www.educativa.com/blog-articulos/gamificacion-el-aprendizaje-divertido/

Gao, J., Cao, Z. y Li, W. (2024). SelfCP: Compressing over-limit prompt via the frozen large language model itself [Cited by: 5; All Open Access; Green Open Access]. *Information Processing and Management*, *61*(6). https://doi.org/10.1016/j.ipm.2024.103873

García, F., Rangel, E. y Mera, A. (2020). Gamificación en la enseñanza de las matemáticas: una revisión sistemática. *Telos: Revista de Estudios Interdisciplinarios en Ciencias Sociales*, *1*, 62-75.

Giraldo Forero, A. F. y Orozco Duque, A. F. (2023). Evolución del procesamiento natural del lenguaje. *TecnoLógicas*, *26*(56), e2687. https://www.researchgate.net/publication/370382559

Gogineni, K., Suvizi, A. y Venkataramani, G. (2025). LLMs on a Budget: System-Level Approaches to Power-Efficient and Scalable Fine-Tuning [Cited by: 5; All Open Access; Gold Open Access; Green Open Access]. *IEEE Open Journal of the Computer Society*, *6*, 987-1000. https://doi.org/10.1109/OJCS.2025.3580498

Goodfellow, I., Bengio, Y. y Courville, A. (2016). *Deep Learning* [http://www.deeplearningbook.org]. MIT Press.

Gu, A. y Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*. https://arxiv.org/abs/2312.00752

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. y Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/pdf/2009.03300

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., ... y Sifre, L. (2022). Training Compute-Optimal Large Language Models. https://arxiv.org/pdf/2203.15556

Hu, J. C., Cavicchioli, R. y Capotondi, A. (2025). Embeddings hidden layers learning for neural network compression [Cited by: 0; All Open Access; Green Open Access; Hybrid Gold Open Access]. *Neural Networks*, *191*. https://doi.org/10.1016/j.neunet.2025.107794

Hussain, M., Chen, C., Anwar, M., Ghorashi, S. A., Ahmed, A., Malik, M. S. A. y Yamin, I. (2025). Adaptive multitask emotion recognition and sentiment analysis using resource-constrained MobileBERT and DistilBERT: an efficient approach for edge devices [Cited by: 0; All Open Access; Gold Open Access; Green Open Access]. *PeerJ Computer Science*, *11*. https://doi.org/10.7717/peerj-cs.3287

Iznaga, Y. S., Rato, L., Salgueiro, P. y León, J. L. (2025). Integrating Large Language Models into Automated Software Testing [Cited by: 0; All Open Access; Gold Open Access; Green Open Access]. *Future Internet*, *17*(10). https://doi.org/10.3390/fi17100476

Jia, N., Liu, T., Chen, J., Zhang, Y. y Han, S. (2025). Task-Agnostic Adaptive Activation Scaling Network for LLMs [Cited by: 0; All Open Access; Gold Open Access; Green Open Access]. *IEEE Access*, *13*, 31774-31784. https://doi.org/10.1109/ACCESS.2025.3542415

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Boudart, E., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Le Scao, T., ... y El Sayed, W. (2024). Mixtral of Experts. https://arxiv.org/abs/2401.04088

Jurafsky, D. y Martin, J. H. (2024). *Speech and Language Processing* (3rd edition draft). Stanford University. https://web.stanford.edu/~jurafsky/slp3/

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. y Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*. https://arxiv.org/pdf/2001.08361

Katharopoulos, A., Vyas, A., Pappas, N. y Fleuret, F. (2020). Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*. https://arxiv.org/abs/2006.16236

Khoboko, P. W., Marivate, V. y Sefara, J. (2025). Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models [Cited by: 10; All Open Access; Gold Open Access; Green Open Access]. *Machine Learning with Applications*, *20*. https://doi.org/10.1016/j.mlwa.2025.100649

Khurana, D., Koli, A., Khatter, K. y Singh, S. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. *ResearchGate*. https://www.researchgate.net/publication/350058919

Kim, J.-H. y Choi, Y.-S. (2025). Lightweight Pre-Trained Korean Language Model Based on Knowledge Distillation and Low-Rank Factorization [Cited by: 2; All Open Access; Gold Open Access; Green Open Access]. *Entropy*, *27*(4). https://doi.org/10.3390/e27040379

Kitchenham, B. y Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering* (inf. téc. N.º EBSE-2007-01). Keele University y University of Durham. https://www.researchgate.net/publication/302924724_Guidelines_for_performing_Systematic_Literature_Reviews_in_Software_Engineering

Kong, F., Shan, X., Hu, Y. y Li, J. (2025). Automated UAV Object Detector Design Using Large Language Model-Guided Architecture Search [Cited by: 2; All Open Access; Gold Open Access; Green Open Access]. *Drones*, *9*(11). https://doi.org/10.3390/drones9110803

Korthikanti, V., Jiali, Y., Shoeybi, M., Patwary, M. et al. (2022). Reducing Activation Recomputation in Large Transformer Models. *arXiv preprint arXiv:2205.05198*. https://arxiv.org/pdf/2205.05198

Kossayakova, A., Ildar, K., Spada, L. L., Zeeshan, N., Bakyt, M., Khuralay, M. y Abdirashev, O. (2025). Large language models for pattern recognition in text data [Cited by: 0; All Open Access; Gold Open Access]. *IAES International Journal of Artificial Intelligence*, *14*(6), 5311-5332. https://doi.org/10.11591/ijai.v14.i6.pp5311-5332

Krajewski, J., Ludziejewski, J., Adamczewski, K., Pióro, M. et al. (2024). Scaling Laws for Fine-Grained Mixture of Experts. *arXiv preprint arXiv:2402.07871*. https://arxiv.org/pdf/2402.07871

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H. y Stoica, I. (2023). Efficient Memory Management for Large Language Model Serving with PagedAttention. *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (SOSP '23)*. https://doi.org/10.1145/3600006.3613165

Kyung, K., Yun, S. y Ahn, J. H. (2025). SSD Offloading for LLM Mixture-of-Experts Weights Considered Harmful in Energy Efficiency [Cited by: 1; All Open Access; Green Open Access]. *IEEE Computer Architecture Letters*, *24*(2), 265-268. https://doi.org/10.1109/LCA.2025.3592563

Leroux, N., Manea, P.-P., Sudarshan, C., Finkbeiner, J., Siegel, S., Strachan, J. P. y Neftci, E. (2025). Analog in-memory computing attention mechanism for fast and energy-efficient large language models [Cited by: 7; All Open Access; Green Open Access; Hybrid Gold Open Access]. *Nature Computational Science*, *5*(9), 813-824. https://doi.org/10.1038/s43588-025-00854-1

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Atkins, M. et al. (2022). Holistic Evaluation of Language Models. *arXiv preprint arXiv:2211.09110*. https://www.researchgate.net/publication/365448636_Holistic_Evaluation_of_Language_Models

Lieber, O., Lenz, B., Bata, H., Cohen, G., Osin, J., Dalmedigos, I., Safahi, E., Meirom, S., Belinkov, Y., Shalev-Shwartz, S., Abend, O., Alon, R., Asida, T., Bergman, A., Glozman, R., Gokhman, M., Manevich, A., Ratner, N., Rozen, N., ... y Shoham, Y. (2024). Jamba: A Hybrid Transformer–Mamba Language Model. https://arxiv.org/abs/2403.19887

Liu, N., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F. y Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, *12*, 157-173. https://doi.org/10.1162/tacl_a_00638

Liu, Z., Zeng, H. y Yang, G. (2025). Research on Lightweight LLM Optimization Methods Based on Knowledge Distillation [Cited by: 0; All Open Access; Gold Open Access]. *Advances in Transdisciplinary Engineering*, *75*, 1935-1942. https://doi.org/10.3233/ATDE250944

Luccioni, A. S., Viguier, S. y Ligozat, A.-L. (2023). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *Journal of Machine Learning Research*, *24*(200), 1-15. https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf

Mahdi, M. A., Fati, S. M., Ragab, M. G., Hazber, M. A. G., Ahamad, S., Saad, S. A. y Al-Shalabi, M. (2025). A Novel Hybrid Attention-Based RoBERTa-BiLSTM Model for Cyberbullying Detection [Cited by: 1; All Open Access; Gold Open Access; Green Open Access]. *Mathematical and Computational Applications*, *30*(4). https://doi.org/10.3390/mca30040091

Martínez, H., Catalán, S., Castelló, A., Mestre, J. I. y Quintana-Orti, E. S. (2025). Latency-Critical Quantized Inference with Transformer Decoders on ARM and RISC-V CPUs. *IEEE Internet of Things Journal*, *12*(13), 25676-25690. https://doi.org/10.1109/JIOT.2025.3560382

Meier, S., Raut, P. N., Mahr, F., Thielen, N., Franke, J. y Risch, F. (2025). Structured knowledge-based causal discovery: Agentic streams of thought [Cited by: 3; All Open Access; Hybrid Gold Open Access]. *Information Processing and Management*, *62*(5). https://doi.org/10.1016/j.ipm.2025.104202

Nair, S., Rao, Y. y Shankarmani, R. (2024). Assessment of Transformer-Based Encoder-Decoder Model for Human-Like Summarization. *arXiv preprint arXiv:2410.16842*. https://arxiv.org/pdf/2410.16842

Ning, W., Shi, S., Yin, Y., Guo, Z. y Xie, J. (2025). Intent-aware personalized summarization for educational texts with large language models [Cited by: 0; All Open Access; Gold Open Access; Green Open Access]. *Journal of King Saud University - Computer and Information Sciences*, *37*(9). https://doi.org/10.1007/s44443-025-00333-7

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M. y Dean, J. (2021). Carbon Emissions and Large Neural Network Training. *arXiv preprint arXiv:2104.10350*. https://arxiv.org/pdf/2104.10350

Pau, D. P. y Aymone, F. M. (2024). Forward Learning of Large Language Models by Consumer Devices [Cited by: 6; All Open Access; Gold Open Access]. *Electronics (Switzerland)*, *13*(2). https://doi.org/10.3390/electronics13020402

Pearce, T. y Song, J. (2024). Reconciling Kaplan and Chinchilla scaling laws.

Pope, R., Douglas, S., Chowdhery, A., DeVries, J., Guy, C., Heek, J., Heilig, O., Hawkins, J. P., Coffer, J. D., Shilitz, M. J. et al. (2023). Efficiently Scaling Transformer Inference. *Proceedings of Machine Learning and Systems (MLSys)*, *5*.

Pramanik, S. y Bin Husin, M. H. (2025). The Retrieval-Augmented Pedagogical Assistant (RAPA): A Methodology for Enhancing Critical Thinking and Equity in AI-Augmented

Education [Cited by: 0; All Open Access; Gold Open Access]. *International Journal of Advanced Computer Science and Applications*, *16*(12), 274-284. https://doi.org/10.14569/IJACSA.2025.0161226

Samsi, S., Zhao, D., McDonald, J., Li, B., Maghazeh, A., Tiwari, D. y Gadepally, V. (2023). From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. *2023 IEEE High Performance Extreme Computing Conference (HPEC)*. https://doi.org/10.1109/HPEC58863.2023.10363447

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*(3), 379-423, 623-656. https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G. y Dean, J. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *International Conference on Learning Representations (ICLR)*. https://www.cs.toronto.edu/~hinton/absps/Outrageously.pdf

Sheik, R., Siva Sundara, K. y Nirmala, S. J. (2024). Neural Data Augmentation for Legal Overruling Task: Small Deep Learning Models vs. Large Language Models [Cited by: 9; All Open Access; Hybrid Gold Open Access]. *Neural Processing Letters*, *56*(2). https://doi.org/10.1007/s11063-024-11574-4

Shen, H. y Hu, B. (2025). Distance-Based Compression Method for Large Language Models [Cited by: 0; All Open Access; Gold Open Access; Green Open Access]. *Applied Sciences (Switzerland)*, *15*(17). https://doi.org/10.3390/app15179482

Stewart, J., Redlin, L. y Watson, S. (2012). *Precálculo matemáticas para el cálculo* (2.ª ed.). Cengage.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W. y Liu, Y. (2024). RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, *568*, 127063. https://arxiv.org/pdf/2104.09864

Udrescu, A. y Popovici, D.-M. (2025). Agora: A Distributed Language Model Framework with API-Call Support for Integrated Climate Forecasting [Cited by: 0; All Open Access; Gold Open Access; Green Open Access]. *IEEE Access*, *13*, 84112-84129. https://doi.org/10.1109/ACCESS.2025.3568028

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. y Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing*

Systems (NeurIPS). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, Z., Luo, T., Liu, C., Liu, W., Goh, R. S. M. y Wong, W.-F. (2025). Enabling Energy-Efficient Deployment of Large Language Models on Memristor Crossbar: A Synergy of Large and Small [Cited by: 5; All Open Access; Green Open Access]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(2), 916-933. https://doi.org/10.1109/TPAMI.2024.3483654

Wang, Z. Q. y El Saddik, A. (2023). DTITD: An Intelligent Insider Threat Detection Framework Based on Digital Twin and Self-Attention Based Deep Learning Models [Cited by: 31; All Open Access; Gold Open Access; Green Open Access]. *IEEE Access*, *11*, 114013-114030. https://doi.org/10.1109/ACCESS.2023.3324371

Wei, J., Wang, X., Schuurmans, D., Maarten, B., Ichter, B., Xia, F., Chi, E., Le, Q. y Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*. https://www.researchgate.net/publication/358232899_Chain_of_Thought_Prompting_Elicits_Reasoning_in_Large_Language_Models

Wen, F., Lu, X., Yu, H., Lu, C., Li, H. y Shi, X. (2025). Dynamic Asymmetric Attention for Enhanced Reasoning and Interpretability in LLMs [Cited by: 0; All Open Access; Gold Open Access]. *Symmetry*, *17*(8). https://doi.org/10.3390/sym17081303

Wu, H., Hong, H., Mao, J., Yin, Z., Wu, Y., Bai, X., Sun, L., Pu, M., Liu, J. y Li, Y. (2025). Forging Robust Cognition Resilience in Large Language Models: The Self-Correction Reflection Paradigm Against Input Perturbations [Cited by: 0; All Open Access; Gold Open Access; Green Open Access]. *Applied Sciences (Switzerland)*, *15*(9). https://doi.org/10.3390/app15095041

Yu, Y.-C., Ouyang, Y.-C. y Lin, C.-A. (2025). TriP-LLM: A Tri-Branch Patch-Wise Large Language Model Framework for Time-Series Anomaly Detection [Cited by: 0; All Open Access; Gold Open Access; Green Open Access]. *IEEE Access*, *13*, 168643-168653. https://doi.org/10.1109/ACCESS.2025.3613663