# Mamba: Modelo SSM como alternativa a los Transformers, nuevas posibilidades en IA.

Cruz Morete, Daniela Elizabeth*, Bestard Columbié, Ana Beatriz **, Soler Jay, Lianet***.
* Estudiante de 2do año de Ingeniería Informática. Universidad de Oriente. Cuba. daniela.cruz@estudiante.uo.edu.cu
** Estudiante de 2do año de Ingeniería Informática. Universidad de Oriente. Cuba. ana.bestard@estudiante.uo.edu.cu
*** Estudiante de 2do año de Ingeniería Informática. Universidad de Oriente. Cuba. lianet.soler@estudiante.uo.edu.cu

**Resumen**
Este artículo aborda el surgimiento de nuevos modelos de lenguaje como alternativa a los Transformers, centrándose en la línea de investigación sobre modelos de secuencia eficientes a partir de modelos de espacio de estado (SSM), y en particular, desde el surgimiento del modelo Mamba, su evolución hasta Mamba2 y el surgimiento de modelos híbridos que utilizan lo mejor de diversas arquitecturas, lo cual ha propiciado el surgimiento entre otros, de los modelos Jamba-1.5 y Mamba2-Hybrid. Se describe las innovaciones en arquitectura, las mejoras en la eficiencia de cómputo y la capacidad de procesamiento de secuencias largas. Como resultado de la investigación se pudo corroborar la existencia de mejoras significativas en el rendimiento y la eficiencia debido a las características de alta velocidad de inferencia, la escalabilidad y los resultados de última generación de Mamba en diferentes tipos de datos, particularmente en el manejo de contextos extensos y las implicaciones de este resultado para el futuro del procesamiento del lenguaje natural y la inteligencia artificial.

**Palabras clave:** modelos de lenguaje, arquitecturas híbridas, Mamba, Jamba-1.5, procesamiento de secuencias largas.

**Abstract**
This article deals the emergence of new language models as an alternative to Transformers, focusing on the line of research on efficient sequence models based on state space models (SSM), and in particular, since the emergence of the Mamba model, its evolution to Mamba2 and the emergence of hybrid models that use the best of various architectures, which has led to the emergence, among others, of the Jamba-1.5 and Mamba2-Hybrid models. Innovations in architecture, improvements in computational efficiency, and long sequence processing capacity are described. As a result of the research, it was possible to corroborate the existence of significant improvements in performance and efficiency due to the high inference speed characteristics, scalability and state-of-the-art results of Mamba in different types of data, particularly in the handling of extensive contexts and the implications of this result for the future of natural language processing and artificial intelligence.

**Keywords:** language models, hybrid architectures, Mamba, Jamba-1.5, long sequence processing.

## 1. Introducción

Desde finales del año 2022 la Inteligencia Artificial (IA) ha irrumpido en la vida de millones de habitantes del planeta y ha pasado a ser de uso cotidiano, lo cual ha sido posible gracias a la acelerada evolución que los modelos de lenguaje han experimentado en los últimos años. A pesar de que los modelos Transformers tradicionales, han sido hasta el momento ampliamente utilizados, la comunidad científica continúa enfocada en lograr mejores resultados que le permita disminuir las limitaciones desde el punto de vista de cómputo, y ser eficientes en el procesamiento de secuencias largas y complejas [1]. Para contrarrestar estas limitaciones, se ha desarrollado nuevas arquitecturas que buscan combinar diferentes enfoques de manera que el rendimiento y la eficiencia se optimicen simultáneamente.

Este artículo es una continuación de la evolución de estos modelos, explora los modelos Mamba, Mamba2, y Jamba-1.5. Se describen los cambios más notables entre estos modelos, junto con mejoras de rendimiento y eficiencia, y el papel que desempeñarán los modelos híbridos en el futuro desarrollo del procesamiento del lenguaje natural.

## 2. Metodología

El informe se basa en un estudio comparativo de la arquitectura, características y rendimiento de los modelos Mamba, Mamba2 y Jamba-1.5. Se revisó la documentación técnica y los artículos de investigación que contiene la evaluación de rendimiento publicadas por los desarrolladores de los modelos, así como por la comunidad científica.

Los criterios de análisis se circunscribieron a:
1. Arquitectura del modelo
2. Complejidad computacional
3. Capacidad de manejo de secuencias largas
4. Rendimiento en tareas de procesamiento del lenguaje natural
5. Eficiencia en el uso de recursos computacionales

## 3. Resultados

### 3.1 Modelo Mamba

El modelo Mamba introduce una nueva arquitectura para el procesamiento de secuencias, que se centra en la eficiencia computacional y el manejo de secuencias largas. [1] Utiliza espacios de estados selectivos en lugar de la atención tradicional. Esto le permite recordar y olvidar selectivamente información, mejorando la eficiencia computacional. El modelo mantiene un estado de tamaño fijo que se actualiza paso a paso a medida que se procesan las entradas, lo que lo hace efectivo para manejar secuencias largas. Utiliza una función de activación no lineal para calcular el nuevo estado en cada paso y utiliza una capa de salida para generar predicciones basadas en el estado actual.

La arquitectura de Mamba se fundamenta en una red neuronal que optimiza la eficiencia computacional mediante el uso combinado de memoria en la GPU. A diferencia de los Transformers, que dependen en gran medida de mecanismos de atención y capas de preenfoque multicapa, Mamba emplea una estructura de red neuronal de extremo a extremo que no incorpora estos elementos. A pesar de esta diferencia, Mamba genera texto con bastante eficacia y logra una mayor eficiencia computacional que los modelos basados en Transformers [2].

Mamba posee una arquitectura híbrida estructurada por capas de SSM intercaladas con capas de redes neuronales feed-forward, a diferencia de los Transformers que están estructurados por capas de atención alternadas con feed-forward. Además, Mamba utiliza la memoria disponible en la GPU de forma inteligente. Esta diferenciación en el uso de la memoria facilita un acceso más rápido a los datos críticos, lo que contribuye significativamente a la eficiencia computacional

Algunas de las características y ventajas observadas fueron:

1. **Complejidad Computacional:** Tiene una complejidad lineal con respecto a la longitud de la secuencia, mientras que los Transformers poseen una complejidad cuadrática [2].
2. **Captura de Dependencias a Largo Plazo**: Introdujo mejoras en la capacidad del modelo para manejar secuencias largas o procesamiento de texto extenso sin perder rendimiento. [3].
3. **Arquitectura:** Presenta una arquitectura simplificada, que le permite lograr entrenamiento de forma más rápida.

### 3.2 Modelo Jamba-1.5

Jamba-1.5 es una evolución de Jamba, siendo este un modelo que combina Mamba con Transformers para arreglar algunos de los inconvenientes de Mamba en cuanto a los niveles de calidad de respuesta. y fue la primera implementación de producción a gran escala de Mamba,

La arquitectura de Jamba conocida como arquitectura SSM-Transformer Jamba [3] se basa en un diseño de capas y bloques que permite una integración efectiva de las arquitecturas Transformer y Mamba. Cada bloque de Jamba contiene una capa de atención o una capa de Mamba, seguida de un perceptrón multicapa [1](MLP), esto significa que hay una capa de Transformador por cada ocho capas en total.[4]

AI21 es la empresa tras Jamba y recientemente ha dado a conocer la familia de modelos abiertos Jamba-1.5, que incluye los modelos conocidos como Jamba-1.5-mini y Jamba-1.5-Large. Los cuales fueron presentados a la comunidad bajo la Licencia de Modelo Abierto Jamba (Jamba Open Model License Agreement). La ventana de contexto de 256K es la más larga entre los modelos abiertos que logra mantener la misma calidad de rendimiento en los límites superiores de su ventana de contexto [3]. En el benchmark RULER, Jamba-1.5-Large, superó a modelos como GPT-4 y LLaMA [3].

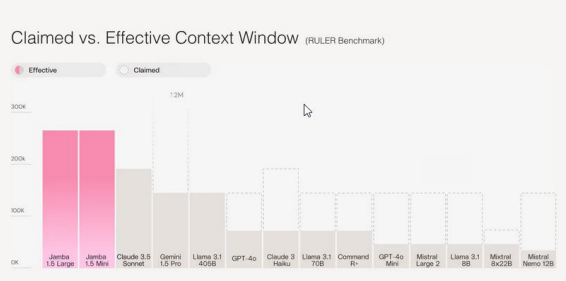


Figura 1. RULER Benchmark [3]

Las características y ventajas observadas fueron:

1. **Arquitectura:** Utiliza una Arquitectura Híbrida avanzada que combina bloques de atención y Mamba (Joint Attention and Mamba) [6].

[1] Red neuronal artificial (RNA) formada por múltiples capas.

2. **Longitud de Contexto Extendida:** Capacidad para manejar un contexto efectivo de hasta 256,000 tokens [4].
3. **Versiones del Modelo**:
- Jamba-1.5-Large: 94 mil millones de parámetros activos
- Jamba-1.5-Mini: 12 mil millones de parámetros activos
4. **Técnica de Cuantización ExpertsInt8**: Permite una inferencia rentable, ejecutando Jamba-1.5-Mini en una sola GPU NVIDIA A100 de 80GB [4].

### 3.3 Modelo Mamba2-Híbrido

Mamba2 es una evolución de Mamba propuesta por sus mismos creadores, que introduce varias mejoras en una arquitectura simplificada, entre ellas el uso de una estructura de identidad de tiempos escalares y que permite dimensiones de estado mucho más grandes, es decir introduce mejoras a su modelo predecesor poniendo el foco de atención sobre las posibilidades que su uso ofrece, a partir de este se ha desarrollado el modelo Mamba2-Hibrido, a cargo de la compañía NVIDIA, este presenta una arquitectura que combina las mejoras de SSM y Transformers y que utiliza la técnica de atención estructurada y enmascarada (SMA) la cual realiza el entrenamiento del modelo de forma más rápida.

1. **Arquitectura simplificada**: La simplificación de su arquitectura permite obtener un mejor rendimiento que el de Mamba.[8]
2. **Eficiencia en Inferencia:** Es de 2 a 8 veces más rápido que Mamba por lo que utiliza menos memoria [5].
3. **Modelos Híbridos:** Combina diferentes tipos de arquitecturas al integrar bloques de MLP, transformadores y SSM, utilizando las mejores características de cada una de estas arquitecturas en función de mejorar aún más su rendimiento. [4]. Posee una velocidad de inferencia hasta 8 veces superior a los transformadores tradicionales en la generación de tokens [6]. Presenta una mayor escalabilidad para tareas de hasta 1 millón de tokens. [7][9].

## 4. Discusión

La evolución de los modelos en los últimos meses muestra una tendencia hacia modelos de lenguaje más eficientes y capaces de manejar contextos extremadamente largos. Cada modelo nuevo ha trabajado sobre las deficiencias ya identificadas del resto de los modelos y se ha enfocado en obtener mejoras, lo cual se traduce en mayor rendimiento y eficiencia de los modelos que trabajaran más rápido y aumentaran su complejidad. La tabla muestra una comparativa entre dos de los últimos modelos híbridos que utilizan Mamba:

Las características y ventajas observadas fueron:

1. **Eficiencia Computacional**: La progresión de complejidad cuadrática a lineal, y las mejoras en la eficiencia de inferencia, han permitido el procesamiento de secuencias más largas con recursos computacionales limitados.

2. **Manejo de Contexto Largo**: El aumento en la capacidad de manejar contextos largos, desde las mejoras iniciales en Mamba hasta los 256,000 tokens en Jamba-1.5, representa un avance significativo para aplicaciones que requieren comprensión de documentos extensos o conversaciones prolongadas.

3. **Arquitecturas Híbridas**: La combinación las arquitecturas de tipo SSM, Transformadores y MLP, logra la optimización del uso de los recursos computacionales en diversas tareas de procesamiento del lenguaje natural, lográndose un mayor rendimiento que al utilizar las tradicionales arquitecturas.
4. **Escalabilidad y Accesibilidad**: El uso de las técnicas de cuantización para reducir el tamaño y optimizar la eficiencia de los modelos, hace que estos modelos avanzados sean más accesibles para los investigadores y desarrolladores que poseen recursos limitados ya que posibilitan la implementación de modelos de aprendizaje profundo en dispositivos de bajas prestaciones de hardware, los cuales tendrán el comportamiento aceptable en cuanto a uso de memoria y velocidad de procesamiento.

| Aspecto | Mamba2-Híbrido | Jamba-1.5 |
|---|---|---|
| **Arquitectura** | Combina MLP, transformadores y modelos de espacio de estado (SSM). Capa de dualidad de espacio de estado estructurado para reducir la complejidad y mejorar la velocidad. | Arquitectura híbrida con elementos de transformadores y Mamba. Enfocado en longitud de contexto, maneja hasta 256,000 tokens. Introduce técnicas de cuantización como ExpertsInt8 para optimizar rendimiento. |
| **Rendimiento** | Velocidad de inferencia hasta 8 veces superior a transformadores tradicionales. Escalabilidad significativa, maneja secuencias de hasta 1 millón de tokens. | Rendimiento sobresaliente en benchmarks académicos y aplicaciones de chatbot. Jamba-1.5-Mini permite ejecución en una sola GPU de 80GB con contextos de 100,000 tokens. |
| **Uso de Recursos** | Reduce la huella de memoria, solo requiere almacenamiento de vectores de tamaño constante. | Cuantización ExpertsInt8 que optimiza el uso de recursos en hardware limitado. |
| **Aplicaciones y Flexibilidad** | Ideal para aplicaciones que requieren procesamiento de lenguaje rápido y eficiente. | Altamente efectivo en tareas que requieren manejo de contexto extenso, como generación de texto y chatbots. |

Tabla 1. Comparación entre modelos híbridos.

## 5. Conclusiones

La evolución de los modelos de lenguaje híbridos, desde Mamba hasta Jamba-1.5, representa un avance significativo para el campo del procesamiento del lenguaje natural. Estos modelos han logrado superar las limitaciones críticas de los transformers tradicionales, ofreciendo una combinación de eficiencia computacional

y capacidad para manejar contextos extremadamente largos.

Las principales contribuciones que se puede atribuir a las investigaciones que dieron origen al modelo Mamba y las mejoras introducidas a partir de su uso y evolución están relacionadas con:

1. La reducción de la complejidad computacional de cuadrática a lineal.
2. La extensión significativa de la longitud de contexto efectiva.
3. La implementación exitosa de arquitecturas híbridas que combinan las fortalezas de diferentes enfoques.
4. La mejora en la accesibilidad de modelos avanzados mediante técnicas de cuantización.

En el futuro las investigaciones relacionadas a estos modelos podrían estar encaminadas a identificar y explorar qué otras combinaciones de arquitecturas pudieran optimizar aún más el rendimiento, así como investigar qué otras aplicaciones específicas pueden beneficiarse de la capacidad de procesamiento de contextos extremadamente largos, iniciando el camino hacia nuevas innovaciones de inteligencia artificial capaces de comprender y generar lenguaje más allá de lo hasta conocido.

**Referencias**


[1] Vaswani, A., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems.

[2] Dao, T., et al. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv preprint arXiv:2312.00752.

[3] AI21 Labs. (2024). Jamba-1.5: A Hybrid Language Model. AI21 Labs Technical Report.

[4] «Towards AI en LinkedIn: JAMBA, the First Powerful Hybrid Model is Here». Accedido: 27 de julio de 2024. [En línea]. Disponible en: https://www.linkedin.com/posts/towards-artificial-intelligence_jamba-the-first-powerful-hybrid-model-is-activity-7186635678888153088-LpYH

[5] T. Dao and A. Gu, "Mamba 2," Hugging Face, 2023. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/mamba2.

[6] Y. Bhaskar, «AI21 Labs Introduces Jamba: The First Production-Grade Hybrid SSM-Transformer Model», Medium. Accedido: 27 de julio de 2024. [En línea]. Disponible en: https://medium.com/@yash9439/ai21-labs-introduces-jamba-the-first-production-grade-hybrid-ssm-transformer-model-7f6f500441a4

[7] NVIDIA. (2024). NeMo Framework User Guide: Mamba Models. NVIDIA Documentation.

[8] Waleffe, R., et al. (2024). An Empirical Study of Mamba-based Language Models. arXiv preprint.

[9] Gu, A., et al. (2022). Efficiently Modeling Long Sequences with Structured State Spaces. International Conference on Learning Representations.