

Inteligencia Artificial: Modelos de Lenguaje

Globales, Open Source y Arquitecturas Alternativas



Modelos Globales



Open Source



Arquitecturas Alternativas

Septiembre 2025

El Panorama Actual de la IA




Diversificación del Ecosistema

- ✓ Más allá de los **Transformers** dominantes
- ✓ **Arquitecturas especializadas** para casos específicos
- ✓ Enfoques diversos para diferentes necesidades




Impacto Global

- ✓ **Modelos globales** ganando relevancia
- ✓ **Open Source** democratizando el acceso
- ✓ Nueva geopolítica de la inteligencia artificial

 Qwen

 Z.ai

 Kimi

 Roboneo

 WAN



Explorar alternativas a los modelos dominantes no es solo una cuestión técnica, sino una necesidad para garantizar un ecosistema de IA **diverso, accesible e innovador**

Modelos de Lenguaje Globales



DeepSeek

Modelo R1

⚡ **Alto rendimiento** en procesamiento de lenguaje

🌐 **Multilingüe** con énfasis en chino

🔧 Arquitectura optimizada para contexto

⚙️ Procesamiento rápido

📄 Contexto amplio



Alibaba

Asistente Quark

🛒 **E-commerce** y atención al cliente

📈 **Análisis** de tendencias y consumidores

🌐 Integración con ecosistema digital

🛒 E-commerce

📈 Análisis de datos



Zhipu AI

AutoGLM Rumination

🔄 **Mecanismo de rumia** para razonamiento

🔄 **Procesamiento iterativo** de respuestas

🔗 Capacidades avanzadas en generación

💡 Razonamiento

🔄 Iterativo



Neuro-Inspirado

Último avance global

🚀 **100x más rápido** que modelos convencionales

🧠 **Inspirado** en estructura cerebral

🔧 Optimizado para hardware especializado

🧠 Neuromórfico

⚡ Alta velocidad

Otros modelos globales destacados:

🤖 Qwen

🧠 Z.ai

👁️ Kimi

⚙️ Roboneo

🏠 WAN

Modelos de Lenguaje Open Source



LLaMA

- ✓ **Comunidad activa** de desarrolladores
- ✓ Versiones desde 7B hasta 65B parámetros
- ✓ Alto rendimiento en tareas de razonamiento

Meta

Versátil



Mistral

- ✓ **Eficiencia** con pocos parámetros
- ✓ Atención con ventana deslizante (SWA)
- ✓ Rendimiento superior en contexto largo

7B

Optimizado



Falcon

- ✓ **Multiquery attention** optimizada
- ✓ Entrenado con 1.5 billones de tokens
- ✓ Destaca en tareas de razonamiento

TII

40B/7B



Bloom

- ✓ **Multilingüe** (46 idiomas)
- ✓ Desarrollado por más de 1000 investigadores
- ✓ Enfoque en accesibilidad y diversidad

BigScience

176B



GPT-NeoX 2.0

- ✓ **Evolución** de GPT-NeoX original
- ✓ Mejoras en estabilidad y entrenamiento
- ✓ Arquitectura paralela optimizada

EleutherAI

20B



Mixtral & Flux

- ✓ **Mixtral**: Modelo de mezcla de expertos
- ✓ **Flux**: Enfoque en generación de imágenes
- ✓ Innovaciones en arquitectura híbrida

8x7B

Híbrido






Free Open Source LLMs



Groq y Servicios de Inferencia Rápida

Groq

Inferencia ultrarrápida con LPU

-  **LPU** - Primer chip diseñado específicamente para inferencia
-  **500 tokens/segundo** - 10x más rápido que GPUs
-  **Eficiencia energética** - Menor consumo que alternativas

25ms

Latencia

10x




Velocidad

\$0.25

Costo/M tokens

OpenRouter

Acceso unificado a múltiples modelos

-  **API unificada** para acceder a diferentes modelos de IA
-  **Enrutamiento inteligente** - Selecciona automáticamente el mejor modelo
-  **Políticas de datos personalizables** - Control total sobre privacidad

25ms

Latencia edge

50+

Modelos

99.9%

Uptime

Casos de Uso y Comparativa

Servicio	Velocidad	Costo (1M tokens)	Casos de uso ideales
Groq	500 tokens/s	\$0.25	Asistentes conversacionales, procesamiento en tiempo real
OpenRouter	Variable por modelo	\$0.20 - \$1.50	Acceso a múltiples modelos, optimización de costos
Replicate	Depende del modelo	\$0.10 - \$2.00	Despliegue de modelos personalizados, experimentación



Replicate

Ejecuta modelos de IA con una API



Anyscale

Plataforma para aplicaciones de IA



Fireworks AI

Inferencia rápida para modelos



Triton

Servidor de inferencia multi-modelo

Productos de Principales Empresas de IA



OpenAI

Líder en modelos de lenguaje

- ✓ **ChatGPT** - Asistente conversacional avanzado
- ✓ **GPT-4o** - Multimodal con audio, texto e imágenes
- ✓ **GPT-5** - Pensamiento integrado y razonamiento

Desde \$20/mes

Visitar



Google Gemini

Modelo multimodal nativo

- ✓ **Gemini 1.5 Pro** - Alto rendimiento en comprensión
- ✓ **Gemini 1.5 Flash** - Optimizado para velocidad
- ✓ **Contexto amplio** - Hasta 1 millón de tokens

Desde \$20/mes

Visitar



Perplexity

Motor de búsqueda inteligente

- ✓ **Respuestas precisas** con fuentes en tiempo real
- ✓ **Pro Search** - Búsqueda profunda y análisis
- ✓ **Citas y referencias** para cada respuesta

Gratis / \$20/mes Pro

Visitar



Microsoft Copilot

Asistente para productividad

- ✓ **Copilot 365** - Integración con Office
- ✓ **Asistencia en tiempo real** para documentos
- ✓ **Seguridad empresarial** y control de datos


Desde \$30/mes

Visitar

Comparativa de Características

Característica	OpenAI	Google	Perplexity	Microsoft
Multimodalidad	Sí	Sí	Parcial	Sí
Contexto amplio	128K	1M	Limitado	Integrado
Acceso a internet	Plus	Sí	Sí	Sí

Modelos Globales y Open Source

 **Qwen**

Modelo multimodal de Alibaba

✓ **Capacidades multimodales** (texto, imagen, audio)

✓ **Contexto amplio** de hasta 1M tokens

✓ **Rendimiento superior** en comprensión y generación

Multilingüe

Alta velocidad

 **DeepSeek**

Modelo R1 de razonamiento avanzado


✓ **Razonamiento complejo** en tareas lógicas

✓ **Arquitectura optimizada** para eficiencia

✓ **Procesamiento** de lenguaje natural avanzado

Resolución problemas

Comprensión

 **Kimi**

Especializado en procesamiento de lenguaje


✓ **Comprensión contextual** profunda

✓ **Procesamiento** de documentos extensos

✓ **Análisis semántico** avanzado

Documentos

Búsqueda

 **Open Source**

LLaMA, Mistral, Falcon y más

✓ **Comunidad activa** de desarrolladores

✓ **Personalización** y adaptabilidad

✓ **Transparencia** en código y datos

Colaborativo

Accesible

Comparativa de Modelos

Modelo	Tamaño	Acceso	Costo	Casos de uso
Qwen	72B - 1.5T parámetros	API + Código abierto	Variable	Procesamiento multimodal, generación de contenido
DeepSeek	7B - 67B parámetros	API + Código abierto	Gratis / \$0.14/M tokens	Razonamiento lógico, resolución de problemas
Kimi	32B - 128B parámetros	API	\$0.05 - \$0.12/M tokens	Análisis de documentos, comprensión contextual
LLaMA/Mistral	7B - 70B parámetros	Código abierto	Gratis (autohospedado)	Personalización, investigación, desarrollo
OpenAI	8B - 1.8T parámetros	API + Modelos pequeños	\$0.15 - \$15/M tokens	Asistentes virtuales, generación de código

Modelo HRM Jerárquico: La Revolución en IA



¿Qué es HRM?

- ✓ **Hierarchical Reasoning Model** - imita estructura cerebral
- ✓ Solo **27 millones** de parámetros
- ✓ Rendimiento **superior** a modelos gigantes



Ventajas

- ✓ **Eficiencia** computacional extrema
- ✓ **Menor costo** de entrenamiento e inferencia
- ✓ **Mejor razonamiento** en tareas complejas



Características Principales

- ✓ **Nivel estratégico** - visión global
- ✓ **Nivel ejecutivo** - implementación detallada
- ✓ **Procesamiento jerárquico** de información



Aplicaciones Prácticas

- ✓ **Razonamiento complejo** y toma de decisiones
- ✓ **Procesamiento** de lenguaje natural avanzado
- ✓ **Sistemas embebidos** con recursos limitados

Estructura Jerárquica del Modelo HRM



Potencial Revolucionario



IA Accesible

Modelos eficientes que funcionan en dispositivos con recursos limitados



Sostenibilidad

Reducción drástica del consumo energético en la computación de IA



Razonamiento Humano

Capacidad para abordar problemas complejos de manera más natural

Algoritmos Alternativos a Transformers



RNN

Redes Neuronales Recurrentes

Procesamiento secuencial

Memoria a corto plazo

Conexiones recurrentes



GRU

Gated Recurrent Unit

Dos puertas

Simplificación de LSTM

Menos parámetros



LSTM

Long Short-Term Memory

Tres puertas

Estado de celda

Memoria a largo plazo



Modelos de Fusión Temporal

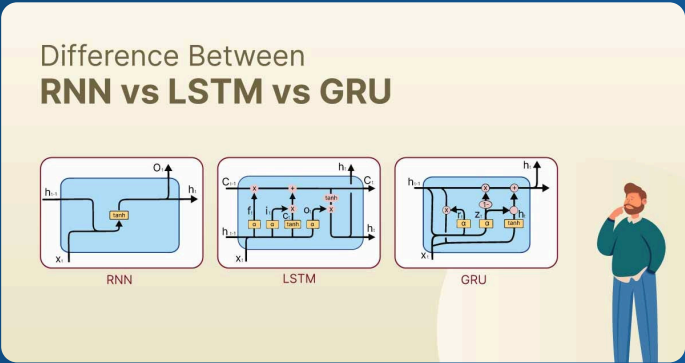
Temporal Fusion Transformers

Combinación de arquitecturas

Atención variable en tiempo

Predicción de series

Ventajas sobre Transformers



Menor Costo Computacional

Requieren menos recursos y energía para entrenamiento e inferencia



Requieren Menos Datos

Efectivos con conjuntos de datos más pequeños y menos diversos



Mejor para Secuencias Cortas

Rendimiento superior en tareas con secuencias de longitud limitada








Procesamiento en Tiempo Real

Menor latencia, ideales para aplicaciones que requieren respuestas inmediatas





Comparativa de Arquitecturas

Ventajas y Desventajas por Arquitectura

Arquitectura	Eficiencia	Rendimiento	Escalabilidad
 Transformers	<div>Alto costo</div> <div>Requiere GPU potente</div>	<div>Excelente</div> en contexto largo <div>Superior</div> en comprensión	<div>Alta</div> <div>Paralelización eficiente</div>
 RNN	<div>Bajo costo</div> <div>Funciona en CPU</div>	<div>Limitado</div> en secuencias largas <div>Problemas de memoria</div>	<div>Baja</div> <div>Dificultad para paralelizar</div>
 LSTM	<div>Moderado</div> <div>Mayor costo que RNN</div>	<div>Bueno</div> en secuencias medias <div>Mejor</div> manejo de memoria	<div>Moderada</div> <div>Mejor que RNN</div>
 GRU	<div>Eficiente</div> <div>Menos parámetros que LSTM</div>	<div>Similar a LSTM</div> <div>Ligeramente inferior</div>	<div>Moderada</div> <div>Mejor paralelización</div>
 HRM	<div>Muy eficiente</div> <div>Solo 27M parámetros</div>	<div>Excelente</div> razonamiento <div>Superior</div> en tareas complejas	<div>Alta</div> <div>Estructura jerárquica</div>







Casos de Uso Óptimos: Transformers

-  Traducción automática de alta calidad
-  Análisis de documentos extensos
-  Asistentes virtuales con contexto amplio
-  Generación de código complejo






Casos de Uso Óptimos: Alternativas

-  Procesamiento en tiempo real con recursos limitados
-  Análisis de series temporales y predicción
-  Reconocimiento de voz en dispositivos móviles
-  Razonamiento complejo con HRM jerárquico

Desafíos y Oportunidades






Desafíos Técnicos

-  **Calidad de datos** en modelos globales
-  **Recursos computacionales** limitados
-  Barreras **lingüísticas** y culturales






Barreras Regulatorias

-  **Geopolítica** de la IA global
-  Restricciones en **transferencia tecnológica**
-  Diferentes **marcos legales** por región






Oportunidades de Innovación

-  **Arquitecturas alternativas** especializadas
-  Optimización para **hardware específico**
-  Enfoques **híbridos** para mayor eficiencia



Colaboración Internacional

-  **Estándares abiertos** para interoperabilidad
-  Comunidades **globales** de investigación
-  Intercambio **académico** sin fronteras

Futuro de la IA más allá de los Transformers



Neuromórfica

Inspirada en estructura cerebral para mayor eficiencia



Auto-optimización

Arquitecturas que se adaptan automáticamente a tareas



Computación Cuántica

Resolución de problemas complejos a velocidad exponencial

Conclusiones



Diversificación del Ecosistema

Modelos globales

Open source

Panorama tecnológico



Arquitecturas Alternativas

RNN/LSTM/GRU

HRM

Casos específicos



Elección Contextual

Recursos

Objetivos

Arquitectura adecuada



Colaboración Global

Estándares abiertos

Cooperación

Futuro de la IA

Visión de Futuro

 Neuromórfica

 Sostenible

 Colaborativa

 Especializada

 Cuántica