Modelos de IA: LLaMA 3

Daniel Canessa Valverde Stephanie Delgado Brenes

Agenda

1. **Introducción a los LLMs**
2. **¿Qué es LLaMA 3?**
3. **Arquitectura y capacidades técnicas**
4. **Rendimiento y benchmarks**
5. **Aplicaciones y casos de uso**
6. **Open source y ecosistema**
7. **Limitaciones y desafíos**
8. **Tendencias y futuro**
9. **Conclusiones**

Introducción al contexto de LLMs

- Evolución de los modelos de lenguaje
  - De GPT a LLaMA: ¿qué cambió?
  - ¿Por qué es relevante hablar de LLaMA 3?
- Importancia del open source en IA generativa

¿Qué es LLaMA 3?

- Modelo de lenguaje grande (LLM) lanzado por Meta en abril de 2024.
- Diseñado para tareas avanzadas de procesamiento de lenguaje natural (como comprensión, generación y razonamiento sobre texto).
- Es open source: pesos y documentación públicos, promoviendo el acceso comunitario (el modelo se puede descargar, modificar y usar).
- LLaMA 3 fue entrenado con más de 15 billones de tokens basados datasets de: common crawl, wikipedia, libros de dominio público, papers científicos y repositorios open source.

Versiones de LLaMA

Modelo	Año	Parámetros (máx.)	Tokens (contexto)	Licencia	Enfoque
LLaMA 1	2023	65B	2,048	Investigación	Democratización inicial
LLaMA 2	2023	70B	4,096	Open source (comercial)	Comunidad, uso comercial
LLaMA 3	2024	8B, 70B	8,192	Open source	Acceso abierto, competitividad
LLaMA 3.1	2024	8B, 70B, 405B	128,000	Open source	Máxima capacidad, contexto largo

Nota: - 128K tokens es similar a 90000 palabras, mientras que 8192 a 6000 palabras.
- Se estima que GPT4 tiene 1.7 trillones de parámetros (aproximadamente 4 veces más que LLaMA 3.1).

Objetivos y filosofía

- Acceso abierto y documentación pública.
- Democratización de la IA.
- Transparencia y ética en el desarrollo de inteligencia artificial.

Arquitectura y capacidades técnicas

- Parámetros y tamaños de modelo (8B, 70B, etc.)
- Innovaciones sobre LLaMA 2
- Entrenamiento: datos, compute, alineamiento
- Comparación rápida con GPT-4, Gemini, Claude 3

Rendimiento y benchmarks

Métricas utilizadas para comparar LLMs:

- **MMLU (Massive Multitask Language Understanding)**
  - Evalúa razonamiento multitarea (ciencias, humanidades, matemáticas, etc.).
  - Estándar para comparar LLMs.
- **GPQA (General Knowledge Questions Advanced)**
  - Mide la habilidad del modelo para responder correctamente preguntas avanzadas de conocimiento general.
- **MATH (Mathematics Benchmark)**
  - Evalúa cálculo, álgebra, geometría y razonamiento matemático.
- **HumanEval**
  - Benchmark de referencia para tareas de programación y generación automática de código.
- **MGSM (Multilingual Grade School Math)**
  - Mide competencia matemática básica y habilidades multilingües.
- **DROP (Discrete Reasoning Over Paragraphs)**
  - Mide comprensión lectora y razonamiento complejo sobre textos largos.

Benchmark sobre distintos LLMs

Benchmark	LLaMA 3.1 400B	GPT-4o	Claude 3 Opus	Gemini Pro 1.5
MMLU (%)	83.7	**88.7**	86.8	**81.9**
GPQA (%)	**24.6**	**53.6**	50.4	35.7
MATH (%)	67.8	**76.6**	68.5	**60.1**
HumanEval (%)	84.1	**90.2**	84.9	**67.0**
MGSM (%)	79.0	**90.5**	88.5	**74.5**

Benchmark	LLaMA 3.1 400B	GPT-4o	Claude 3 Opus	Gemini Pro 1.5
DROP (f1)	83.5	**93.4**	86.0	**78.9**

Fuentes: Meta AI Blog, OpenAI Hello GPT-4o, Anthropic Claude 3, Google Gemini Pro, Hugging Face Leaderboard - Julio 2024.

Nota: LLaMA 3.1 supera a modelos open source previos y queda cerca de los modelos comerciales, pero todavía detrás de GPT-4o y Claude 3 en varias pruebas.

Aplicaciones y casos de uso

- Chatbots, asistentes, generación de código
- Integraciones en productos reales (ejemplos)
- Impacto en industria y comunidad open source

Open Source

- **Licencia abierta:**
  - Open source, con pesos y documentación accesible.
  - Permite uso académico, empresarial y personal sin costo.
  - **Restricciones:**
    - No se puede usar LLaMA para crear otro modelo de gran escala sin permiso de Meta.
    - Prohibido utilizarlo para actividades ilícitas.
- **Accesibilidad de los pesos:**
  - Cualquier persona puede descargar los modelos, auditarlos, adaptarlos y hacer fine-tuning localmente o en la nube.
- **Contribuciones de la comunidad:**
  - Modelos multilingües y adaptados para derecho, medicina, programación, etc. Ejemplo: LLaMA-3-8B-Orca.
  - llama.cpp: librería open source que permite correr LLaMA 3.1 en CPU, Raspberry Pi, Macbooks y servidores, sin requerir GPU. Permite reducir el tamaño sin perder mucha precisión y ejecución en dispositivos móviles.

Ecosistema

- **Versatilidad de despliegue:**
  - LLaMA puede ejecutarse en una amplia variedad de plataformas, desde laptops y servidores hasta dispositivos edge y cloud pública.
- **Local:**
  - PC, Mac, servidores (con o sin GPU).
  - Raspberry Pi y dispositivos ARM.
  - Móviles Android (experimental).
- **Cloud y servicios gestionados:**
  - AWS SageMaker
  - Google Vertex AI
  - Azure Machine Learning
- **Contenedores y MLOps:**
  - Fácil integración en pipelines de CI/CD mediante Docker y Kubernetes.
- **Frameworks de soporte:**
  - llama.cpp: inferencia eficiente en CPU/edge/móvil
  - vLLM: inferencia ultra-rápida en GPU
  - Ollama: despliegue y manejo fácil de modelos en local

Limitaciones y desafíos

- Ética, seguridad, sesgos, alucinaciones
- Longitud de contexto y límites técnicos
- Uso responsable y retos de escalabilidad

Tendencias y futuro

Tendencias actuales de LLaMA

- **LLaMA 3.1 (2024):**
  - Modelos de hasta 405B parámetros, ventana de contexto de 128k tokens.
  - Solo texto. Enfoque en rendimiento, multilingüismo y open source.
- **LLaMA 3.2 (2024):**
  - Modelos de 1B, 3B, 11B, 90B parámetros. Ventana de contexto de 128k tokens.
  - **Primera versión multimodal:** modelos especializados en texto y modelos con capacidades de visión (texto + imagen).
  - Incluye variantes ligeras para dispositivos edge (1B, 3B) y modelos grandes para cloud (11B, 90B).
- **LLaMA 3.3 (2024):**
  - Modelo de 70B parámetros, 128k tokens.
  - Enfoque instruccional: mejoras en razonamiento, tareas de programación y multilingüismo.
  - No es multimodal, pero optimizado para rendimiento en tareas complejas.
- **LLaMA 4 (2025):**
  - Modelos "Scout" y "Maverick": 17B parámetros activos, hasta 400B totales (Mixture-of-Experts).
  - Ventana de contexto de hasta **10 millones de tokens**.
  - **Multimodal nativo:** arquitectura optimizada para fusionar modalidades.

Futuro:

- **Modelos aún más grandes y eficientes** (por ejemplo, LLaMA 4 Behemoth en desarrollo).
- **Multimodalidad avanzada:** integración de audio, video y otras modalidades junto a texto e imagen.

- **Ventanas de contexto ultra-largas:** millones de tokens, memoria dinámica y mejores técnicas de manejo de contexto relevante.
- **Personalización y fine-tuning local**: métodos más fáciles y económicos para adaptar los modelos a tareas o dominios específicos.

Conclusiones

- Impacto de LLaMA 3 en el ecosistema de IA
- Democratización de acceso a IA avanzada
- Reflexión final y preguntas

Fuentes

- Meta AI Blog
- Number of Parameters in GPT-4
- OpenAI Hello GPT-4o
- Anthropic Claude 3.5 Sonnet

- Google Gemini AI

- Hugging Face Open LLM Leaderboard (https://ai.meta.com/resources/models-and-libraries/llama-downloads/)).