Modelos de IA: LLaMA 3, características y aplicaciones

Daniel Canessa Valverde Stephanie Delgado Brenes

Agenda

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).

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.

Nota: LLaMA 3.1 supera ampliamente 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 y ecosistema

- Importancia de la licencia y acceso a pesos
- Fine-tuning y extensiones (comunidad)
- Herramientas compatibles: transformers, LLaMA.cpp, Hugging Face, etc.

Limitaciones y desafíos

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

Tendencias y futuro

- Próximos releases: ¿qué esperar?
- Multimodalidad, agentes, eficiencia, interpretabilidad
- Futuro de la IA open source

Conclusiones

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

Referencias

Meta AI Blog Number of Parameters in GPT-4 LLMs benchmark