Introducción al uso de LLMs

Clase VII - Introducción al Aprendizaje Profundo

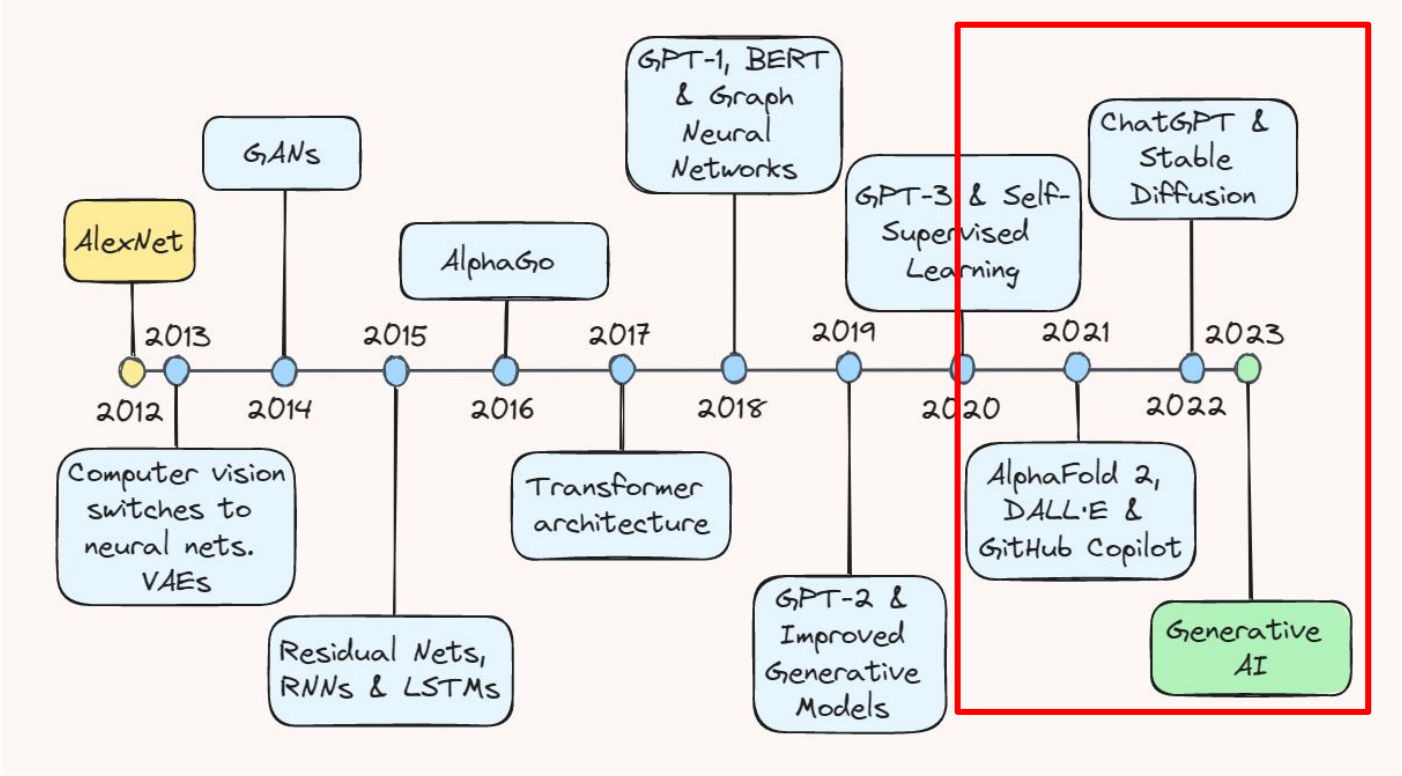
¿Qué vimos hasta ahora?

- Redes neuronales MLP.
- Redes convolucionales.
- Redes recurrentes.
- Redes transformers.

¿Qué vamos a ver hoy?

Mini-introducción al estado de los LLMs

Usted está aquí



Actualidad

- En Nov 2022, OpenAI patea el tablero con ChatGPT.
- **LLM como producto**, comercial y mediáticamente exitoso
- “AI alcance de todos” – puede usarse sin abrir un sólo notebook.
- “AGI feelings” – podemos mantener una “conversación” con ida y vuelta.

Actualidad

- Nueva era de LLMs masivos como producto
- A diferencia de otros hitos de la IA, la mayoría de los LLMs son comerciales, cerrados y con escasos detalles técnicos o metodológicos, poco claro el origen de la información

Actualidad

- Entramos en la era de “Models as a service”, “Pay as you Go”
- Se consumen a través de una API de un proveedor
  - Modelos muy robustos
  - Entrenados con una enorme cantidad de datos de internet
- Varios competidores fuertes
  - OpenAI
  - Meta (Llama, open source models*)
  - Anthropic (Claude)
  - Google (Gemini)
- En esta web puede verse un ranking y comparativa

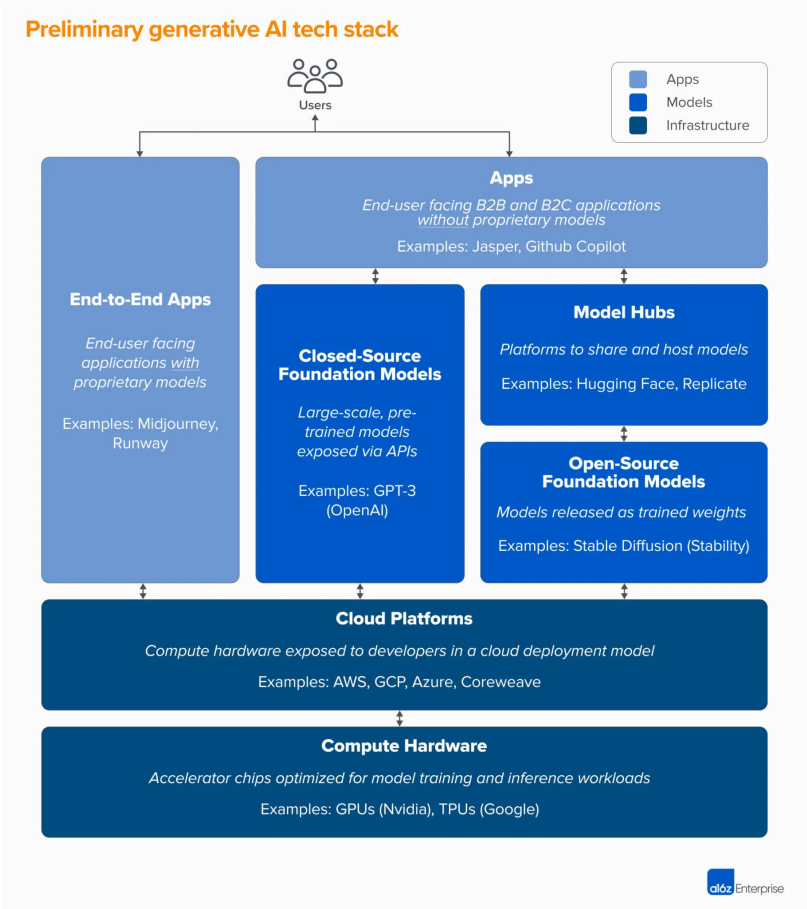
Actualidad

- Competencia feroz entre providers, que sacan constantemente nuevos modelos con mejor capability y a menor precio
- Modelos de propósito general muy buenos
- También de propósito particular como generadores de código o razonadores
- En general se paga por tokens de entrada (contexto) y de salida (generados)

Donde transformers se hacen LLMs...

Actualidad

Ejemplo de aplicación con IA generativa...

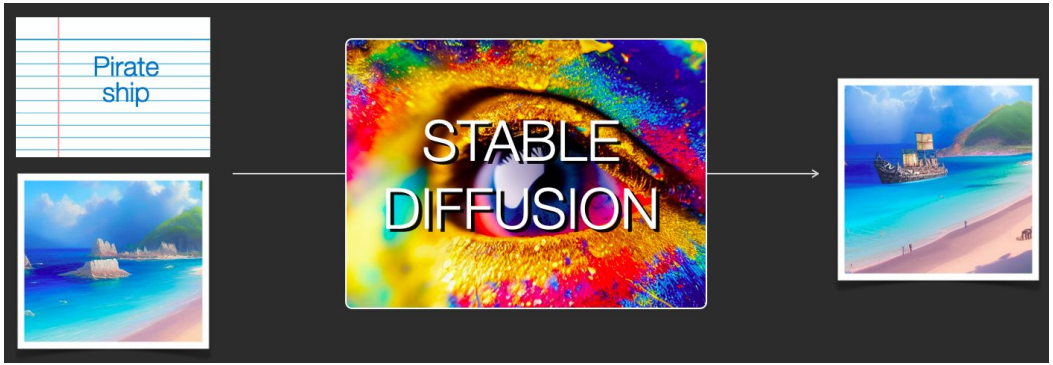


[Fuente: A16z]

Actualidad: modelos multimodales

DALL·E 3 · An expressive oil painting of a chocolate chip cookie being dipped in a glass of milk, depicted as an explosion of flavors.

texto → imágen



imágen, texto → imágen

Algunas técnicas y usos comunes

Algunas técnicas populares

- Prompt engineering – te lleva **MUY** lejos

“Eres un asistente científico experto en crear resúmenes concisos y precisos de artículos de investigación. Tu tarea es resumir el siguiente abstract de un artículo científico en no más de 3 frases.

El resumen debe incluir: 1. El objetivo principal del estudio 2. La metodología utilizada 3. Los hallazgos o conclusiones clave Asegúrate de mantener la terminología científica esencial y expresar las ideas de manera clara y concisa. Evita incluir información secundaria o ejemplos específicos. Abstract a resumir:

[texto del abstract]”

Algunas técnicas populares

- Few shot – agrega ejemplos a la prompt

Tarea: Clasificar la reseña de una película en uno de los siguientes géneros:

Acción, Comedia, Drama, Ciencia Ficción.

Ejemplos: Reseña: "Una película llena de explosiones y persecuciones en coches que te mantiene al borde del asiento." Género: Acción

Reseña: "Me reí de principio a fin. Los diálogos ingeniosos y las situaciones absurdas hicieron que fuera una experiencia hilarante." Género: Comedia

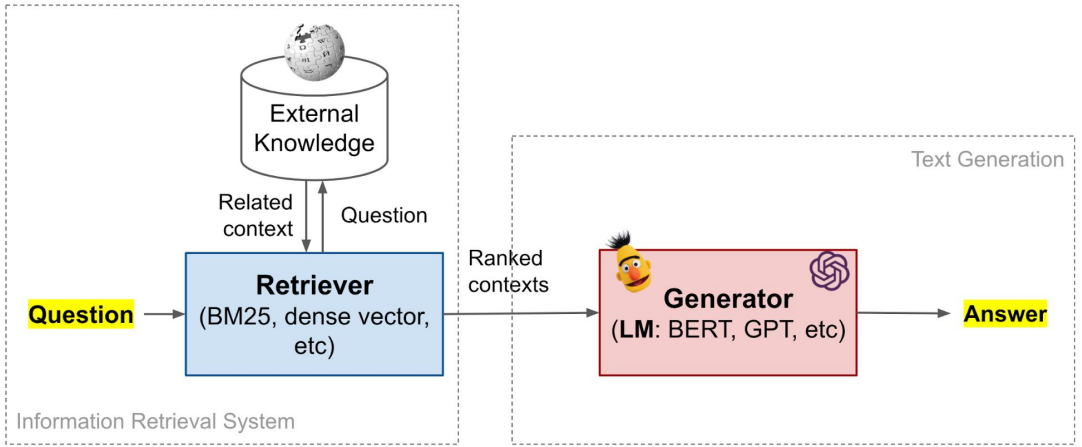
Reseña: "Una historia conmovedora sobre las relaciones familiares y el perdón. Los actores ofrecieron actuaciones emotivas y profundas." Género: Drama

Ahora, clasifica la siguiente reseña: Reseña: "Una exploración intensa de las consecuencias de las decisiones del pasado. Los personajes luchan con sus demonios internos en un viaje emocional que te deja pensando mucho después de que termina la película." Género: [Completar]

Algunas técnicas populares

- Retrieval augmentation generation – para cuando el contexto necesario se hace grande.

  Ej.: usuario pregunta “cuál es la política de equipaje de una aerolínea”
  Como la política total es muy grande, es mejor que el modelo busque en el fragmento relevante



Fuente: Lilian Weng

Algunas técnicas populares

Fine tuning en LLMs...

- Los LLMs ya incluyen fine tuning por ejemplo para búsqueda, coding etc.
- El fine tuning que nosotros queramos hacerle cubre casos muy particulares, en general de dominios muy específicos considerablemente distintos a los que vio en entrenamiento

  “que el modelo hable como cordobés”
  “quiero que el modelo pueda escribir en un lenguaje propietario”

Algunas técnicas populares: Agentes

- Al momento de resolver un problema de negocio, en lugar de usar un único LLM con un prompt grande, podemos dividirlo en subproblemas bien simples y manejables

Nuestras recomendaciones

- Busquen sacarle el máximo provecho posible!!! Ej. como:
  - Editores de código
  - Asistentes personales
  - Sintetizadores de info
  - Generadores de contenido, imágenes, etc.
- LLMs puede acelerar enormemente la velocidad de PoCs
  - Arrancar código asistido con IA
  - Validar viabilidad muy rápidamente antes de tener todo el pipeline
  - Etiquetado de datos
  - Muchos etc.
- Se viene un cambio de paradigma donde la IA va a ser complementaria a muchos de nuestros trabajos

Downsides y aspectos a tener en cuenta

Alucinaciones

- Los LLMs están entrenados para devolver tokens con alta probabilidad
- Esto puede hacer que se den respuestas incorrectas que suenan muy bien
- Esto **es** un problema de cara a lanzar estos modelos a producción

Algunos riesgos de seguridad

- Prompt injection
- Prompt leaking (filtrado de información)

Algunas mitigaciones

- Proveer la respuesta como parte del contexto
- Proveer un esquema que estructure las salidas; si no se cumple, no se considera
- Especificar en el prompt qué hacer cuando no sabe la respuesta
- Implementar funciones tipo guardrails

Ética

- Los LLMs son entrenados usando muy diversas fuentes de datos, que incluyen redes sociales, foros de discusión, etc.
- Por tanto, son propensos a amplificar sesgos y estereotipos con los que fueron entrenados y con los que se les hizo ajuste fino.
- Esto es súper importante de considerar e implementar mitigaciones ej. a la hora de desplegarlos de cara al usuario.

Algunos recursos

- Prompt Engineering Guide
- Courses - DeepLearning.AI
- LLM Bootcamp - The Full Stack
- Hugging Face Open-Source AI Cookbook
- OpenAI Cookbook