

LLM

octubre 2023

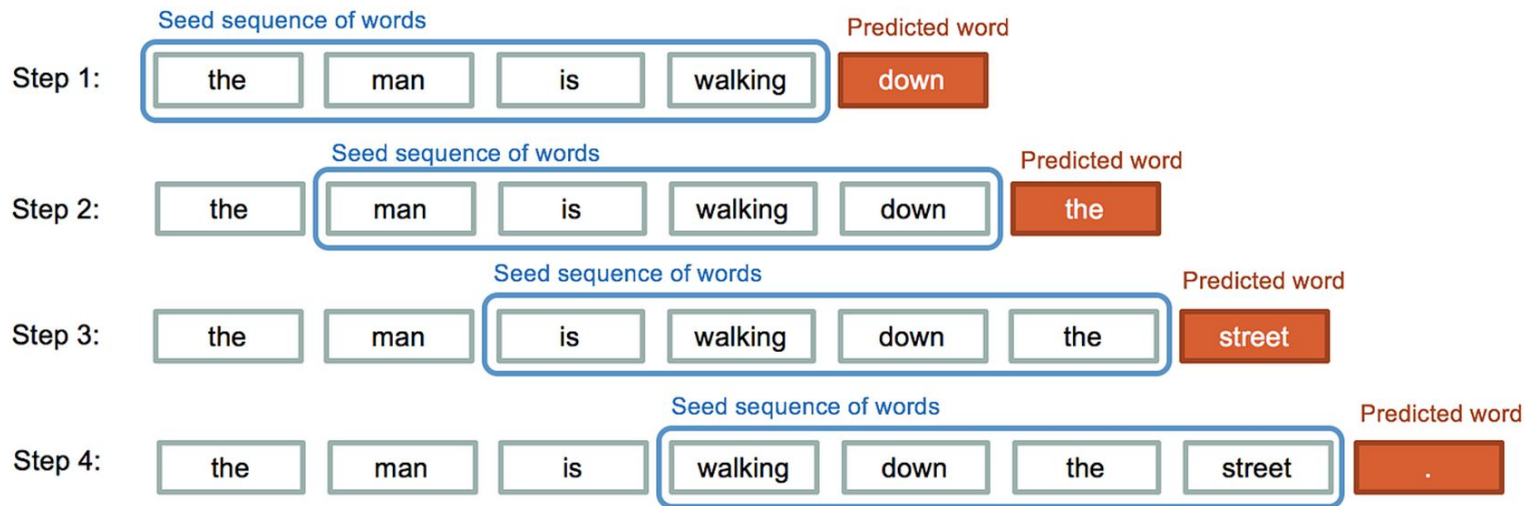
diverger

Predecir la siguiente palabra

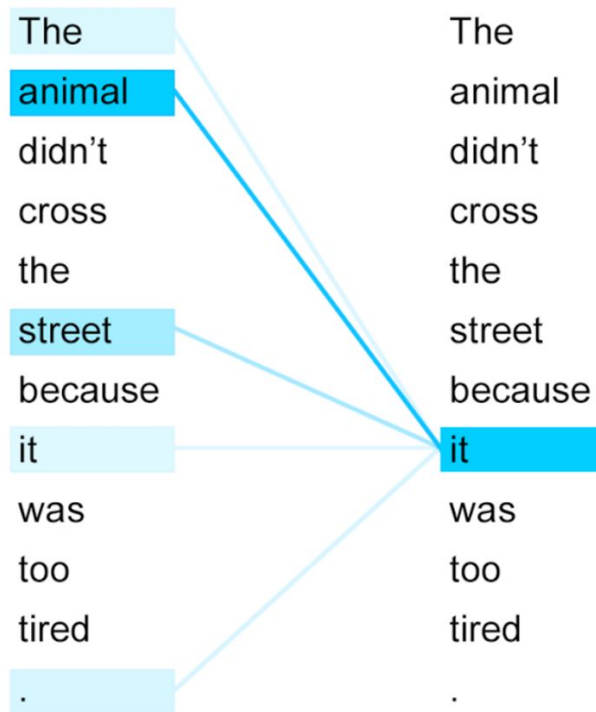
The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

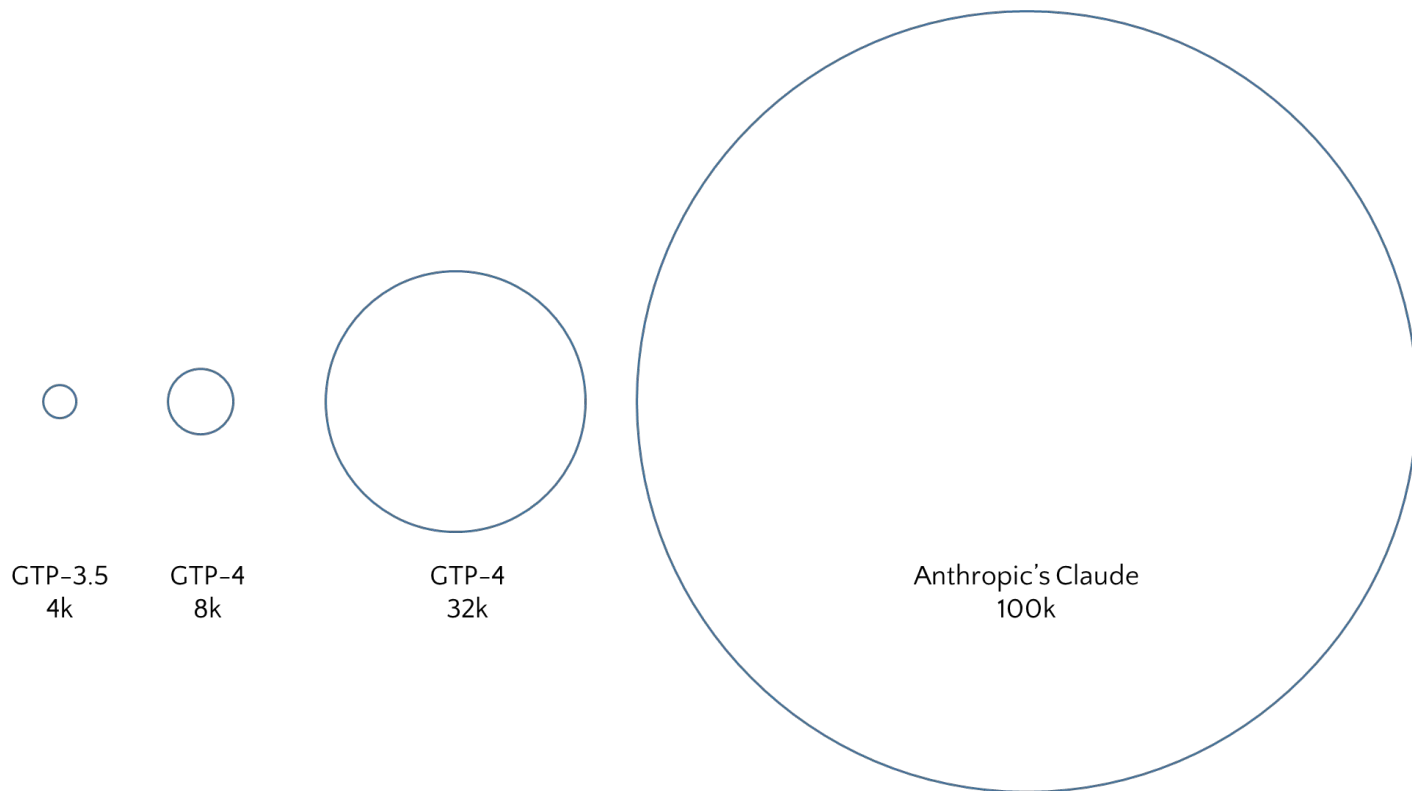
Dentro de un contexto



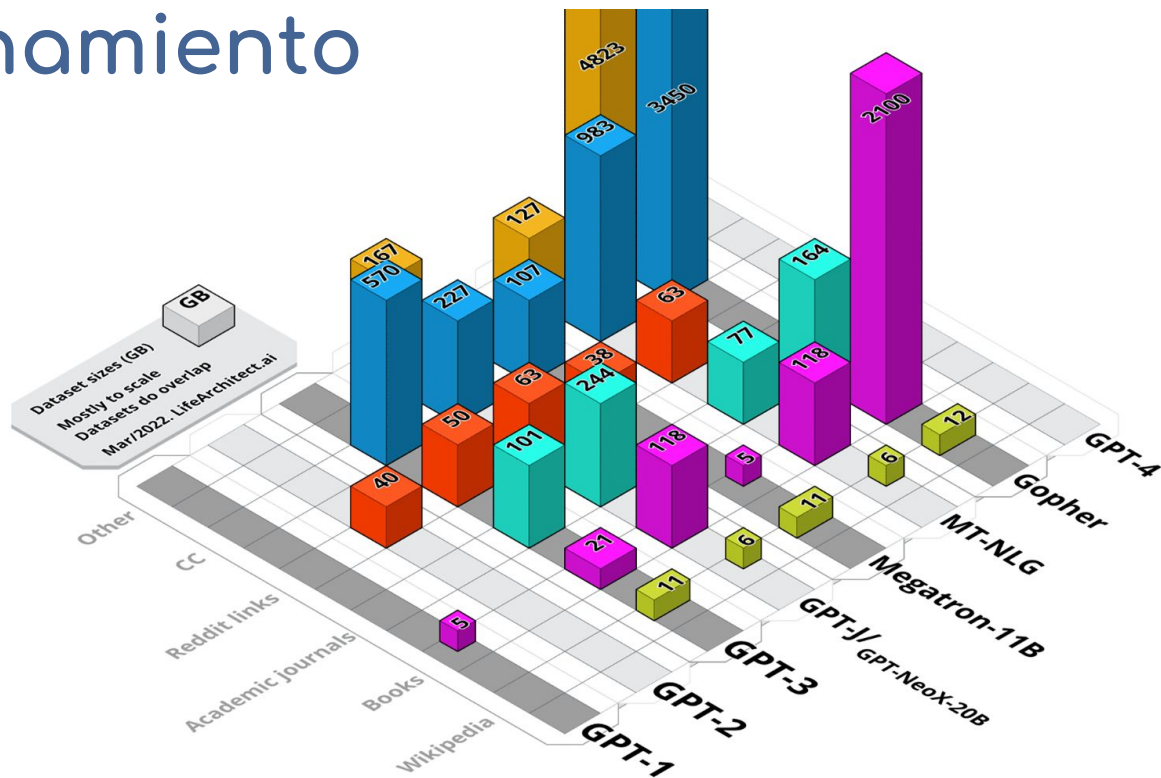
Con atención a lo más relevante



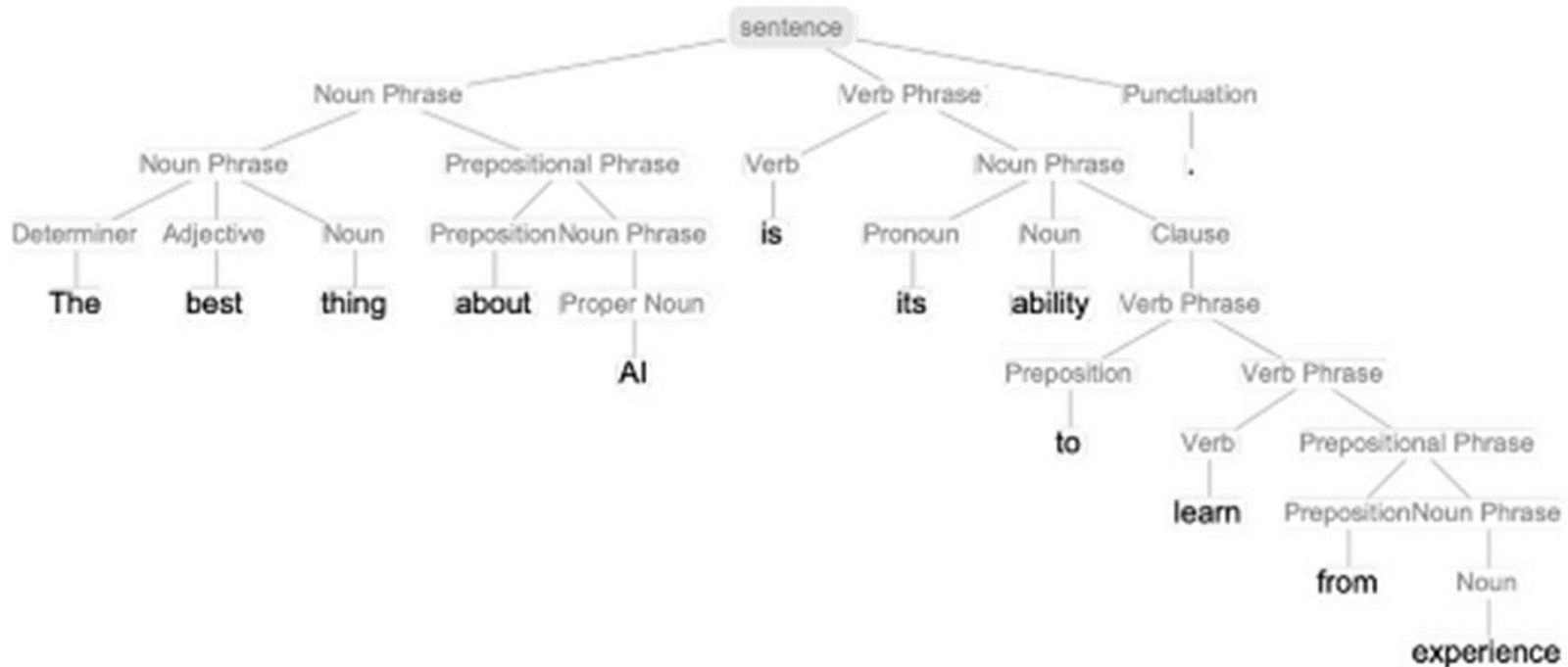
Contextos cada vez más amplios



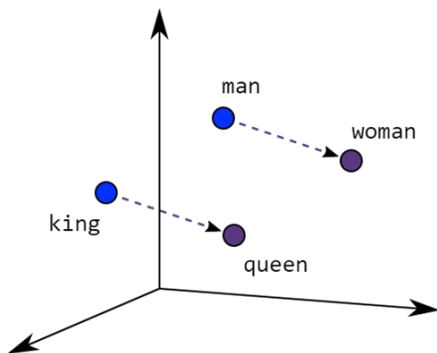
Ingente cantidad de **datos** para el entrenamiento



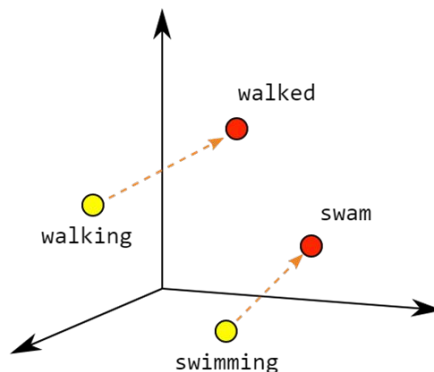
Infiere reglas del lenguaje



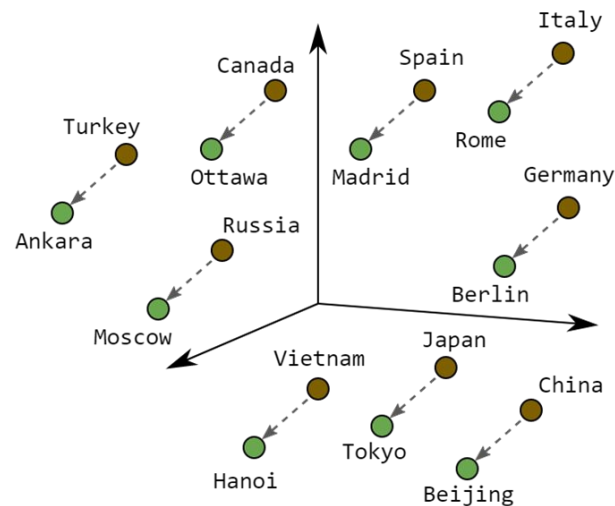
De las palabras a los conceptos



Male-Female

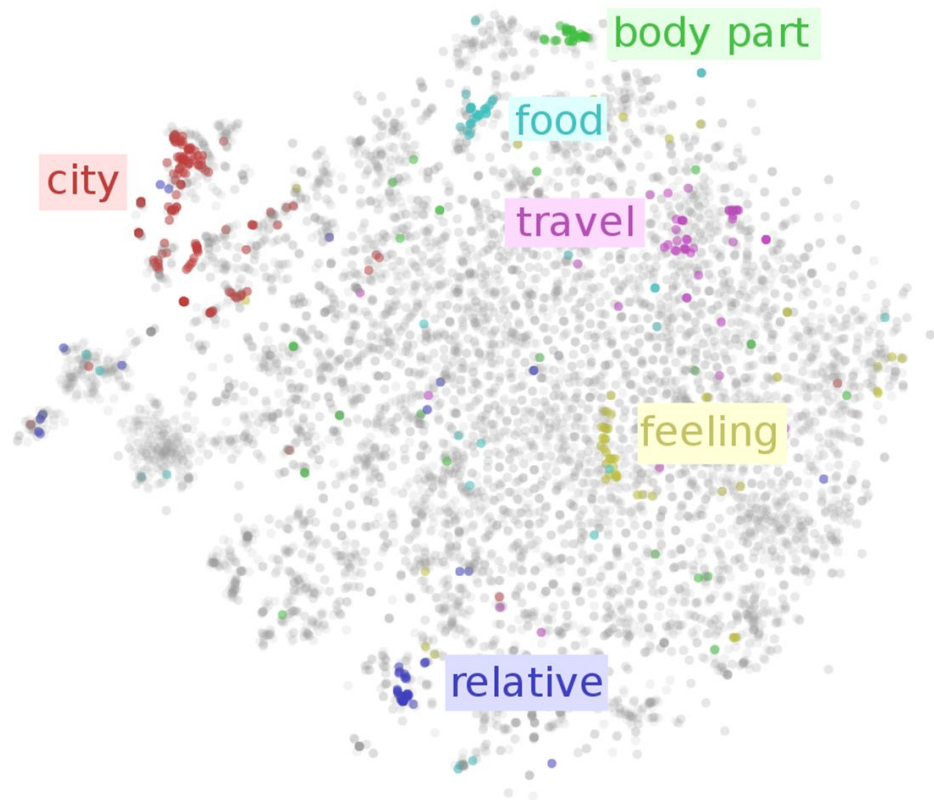


Verb Tense



Country-Capital

¿Un **modelo** conceptual del mundo?



Tipos de actividades basadas en lenguaje

Procesar información

- Analizar pliego
- Analizar comentarios
- Revisiones de calidad

Generar contenido

- Crear informe periódico
- Responder usuarios
- Crear descripciones de producto

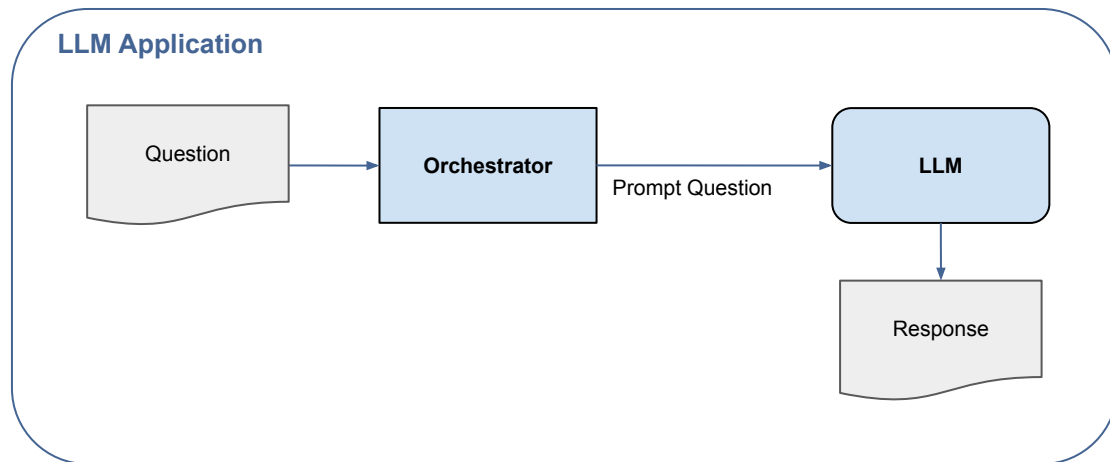
Transformar contenido

- Traducir
- Adaptar contenido en distintos formatos y tonos

Estrategias

Prompts

Question:
Esta es la entrada principal, que representa una consulta que se desea responder utilizando el LLM.



Orchestrator: Convierte la entrada en un prompt, que es una versión mejorada o más detallada de la entrada inicial. Envía el prompt a un LLM y recibe la respuesta.

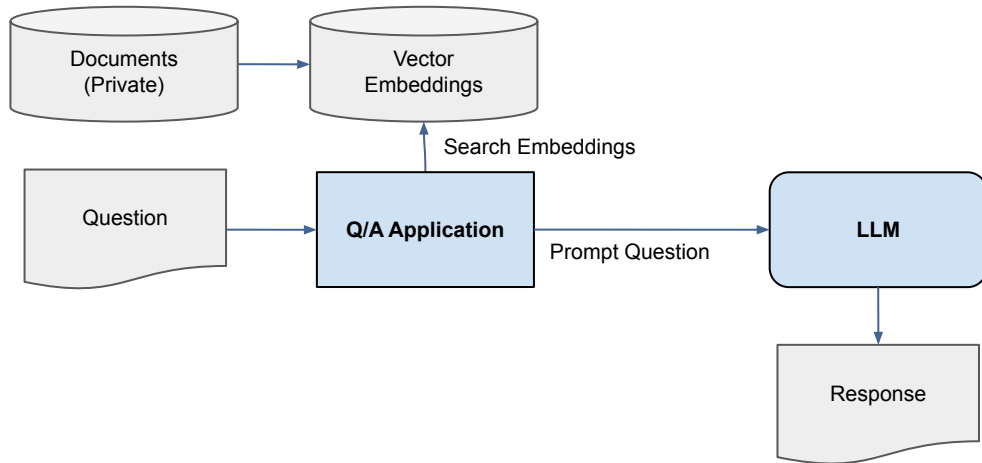
Estrategias

Embeddings

Embeddings: Se crean vectores semánticos (embeddings) a partir del conjunto de documentos y se almacenan en una BBDD especializada.

Question:
Esta es la entrada principal, que representa una consulta que se desea responder utilizando el conjunto de documentos privados de los que se dispone

Retrieval-Augmented Generation (RAG) Method

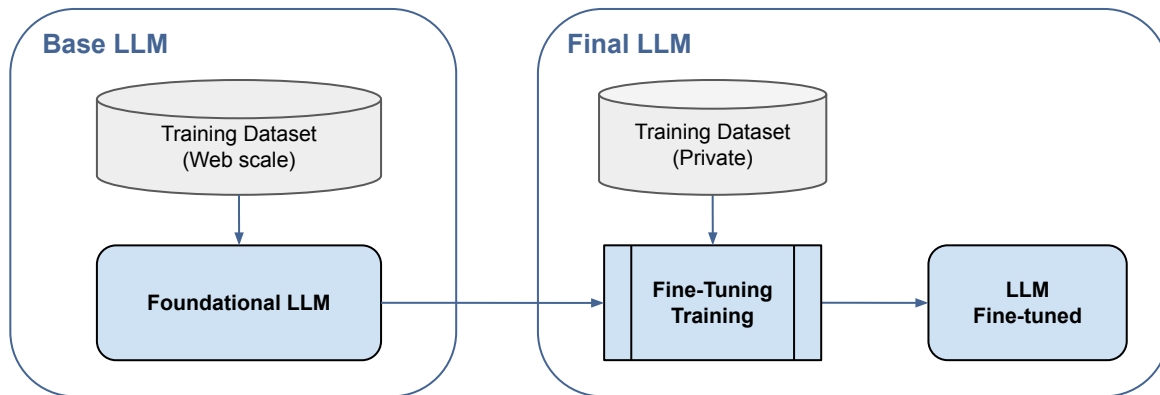


Q/A Application: Procesa la pregunta, obteniendo sus embeddings que utiliza para buscar y recuperar información relevante del conjunto de documentos privados. Con este extracto y la pregunta original se construye un prompt que se envía al LLM para obtener una respuesta detallada.

Estrategias

Fine-Tuning

Base LLM:
Se parte de un modelo fundacional, como GPT-4, PaLM o LLaMA 2, que ha sido pre-entrenado con una gran cantidad de datos.



Fine-tuning: Se realiza otro proceso de entrenamiento sobre el modelo fundacional con otro conjunto de datos, más reducido.

Final LLM: Se obtiene un nuevo modelo más especializado que el fundacional.

Riesgos

Alucinaciones

no crees que Magic Johnson fue mejor que Michael Jordan?

✓ Buscando: 'Magic Johnson vs Michael Jordan'

✓ Generando respuestas para usted...

Esa es una pregunta difícil de responder, ya que Michael Jordan y Magic Johnson son dos de los mejores jugadores de la historia y tienen estilos de juego diferentes ¹. Michael Jordan es el máximo anotador de la NBA y tiene seis campeonatos, seis premios al MVP de las Finales y cinco premios al MVP ². Magic Johnson es el mejor asistente de la NBA y tiene cinco campeonatos, tres premios al MVP y tres premios al MVP de las Finales ². Ambos se enfrentaron en dos Finales de la NBA, en 1991 y 1992, y Jordan ganó las dos ³.

En este contexto, un humano podría decir que Michael Jordan fue mejor que Magic Johnson porque tiene más títulos individuales y colectivos, y porque dominó el baloncesto durante más tiempo. También podría decir que Magic Johnson fue mejor que Michael Jordan porque tenía una visión de juego única, una capacidad para hacer mejores a sus compañeros y una versatilidad para jugar en varias posiciones. Sin embargo, estas son opiniones subjetivas y no representan necesariamente los hechos o la verdad.

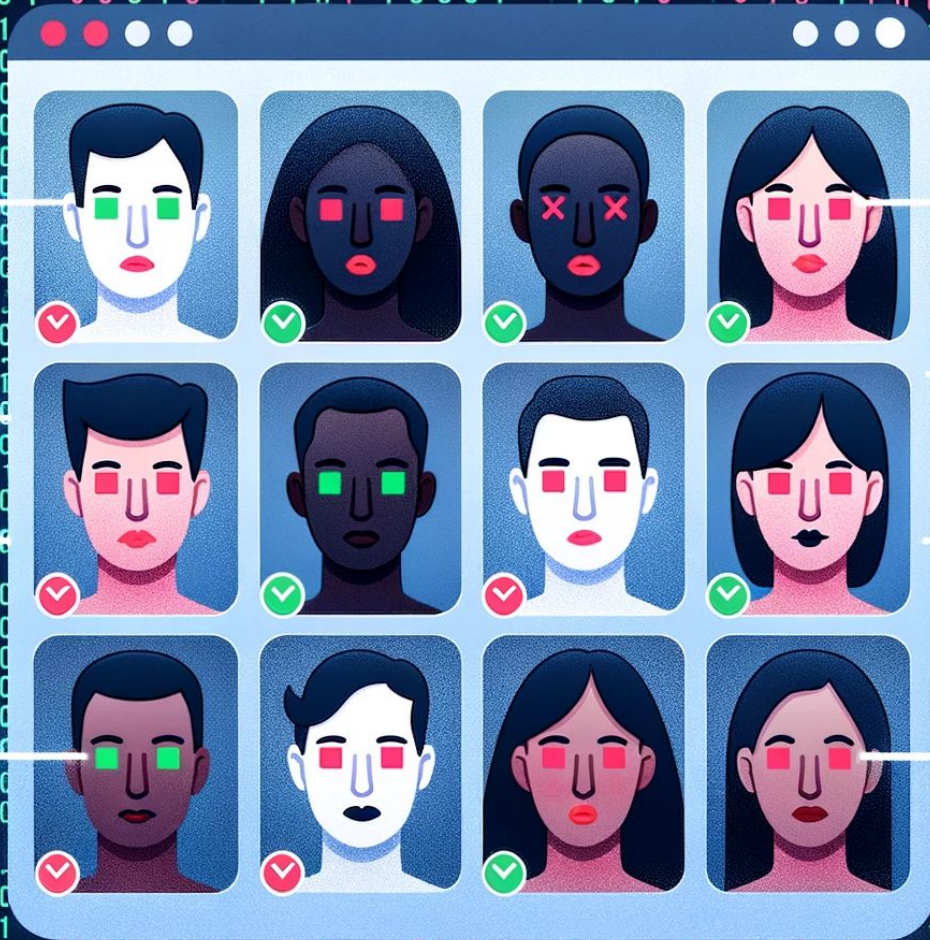
Riesgos

Sesgos



BISAS

ERROR



APPRER

DISCRMINON



Riesgos

Privacidad



Riesgos

Copyright



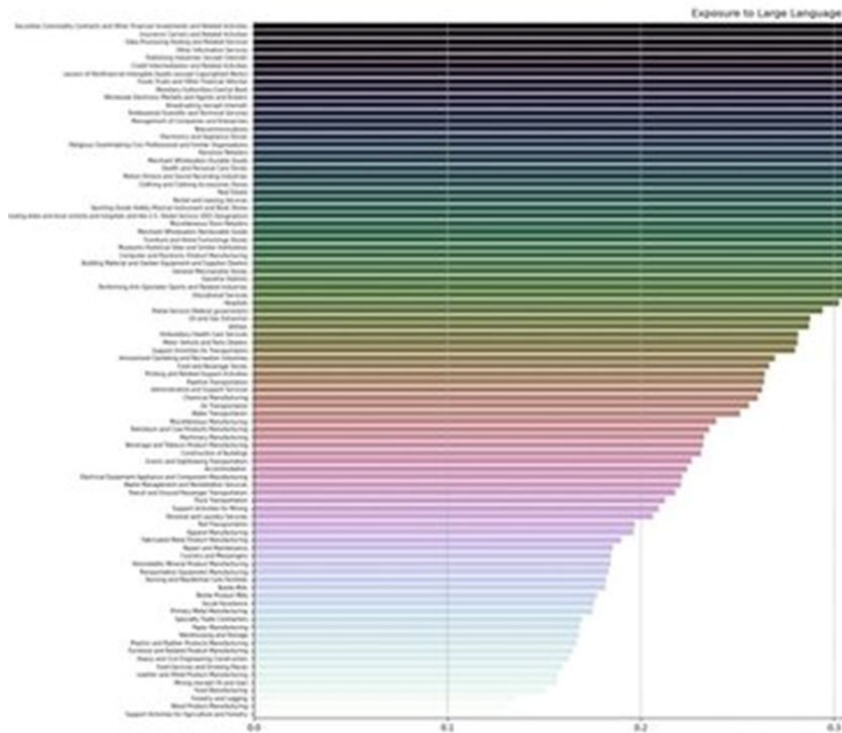
Impacto de los LLM en el mercado laboral

Considerando las capacidades actuales del modelo GPT:

El 19% de los trabajos tienen más del 50% de sus tareas expuestas a los GPTs.

Las ocupaciones de salarios más altos generalmente presentan más tareas con alta exposición.

Las ocupaciones que dependen de habilidades de programación y lenguaje son más susceptibles a ser influenciadas por las LLMs.



“No habrá
programadores
en 5 años.”¹



Emad Mostaque
Founder & CEO Stability AI

¹https://www.youtube.com/watch?v=ciX_iFGyS0M

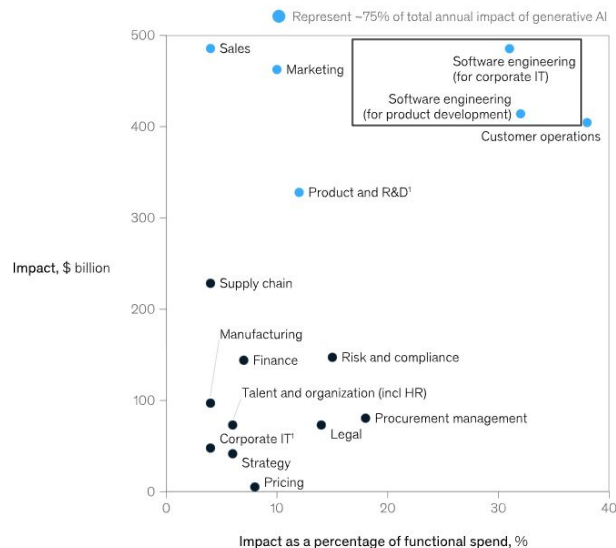
Impacto en el desarrollo de software

Los desarrolladores de software que usaron GitHub Copilot de Microsoft completaron tareas un **56 por ciento** más rápido que aquellos que no utilizaron la herramienta.¹

En 2023 el **10%** del código y las pruebas mundiales será escrito por Inteligencia Artificial.²

El impacto directo de la IA en la productividad de la ingeniería de software podría variar entre el **20 y el 45 por ciento** del gasto anual actual (\$899B).³

Google también está en “el camino para reinventar lo que significa hacer desarrollo de software en la era de los modelos generativos”.⁴



Note: Impact is averaged.

¹Excluding software engineering.

²Source: Comparative Industry Service (CIS), IHS Markit; Oxford Economics; McKinsey Corporate and Business Functions database; McKinsey Manufacturing and Supply Chain 360; McKinsey Sales Navigator; Ignite, a McKinsey database; McKinsey analysis

¹<https://arxiv.org/abs/2302.06590>

²<https://www.forrester.com/report/predictions-2023-artificial-intelligence/RES178186>

³<https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>

⁴<https://twitter.com/DynamicWebPaige/status/1674920912565723136?t=te5THaow6AZJPv-8uAFZhw&s=35>

Tipos de actividades basadas en código

Procesar información

- Explicar código fuente
- Generar comentarios
- Revisiones de calidad
- Identificar errores

Generar contenido

- Scaffolding inteligente
- Crear pruebas unitarias
- Corregir errores

Transformar contenido

- Refactorizar
- Aplicar guías de estilo
- Modernizar aplicaciones

Nuevos desarrollos

Mantenimiento

- Deuda técnica
- Sistemas Legacy

Niveles de adopción

La mayoría de las organizaciones o no han empezado a adoptar masivamente la IA generativa (Piloto) o se encuentran en un uso individual (Copiloto). Desde Diverger pretendemos situar a las organizaciones en estadios de mayor adopción, como son el de torre de control y piloto automático.



Piloto

Los profesionales de su equipo son responsables de todo el ciclo de trabajo, utilizando herramientas convencionales pero sin asistencia directa de IA.



Copiloto

Los profesionales cuentan con un asistente de IA que sugiere opciones, ayuda a completar tareas y a depurar errores, pero aún se mantiene el control del proceso.



Torre de control

Las organizaciones estandarizan la forma en que se utiliza la IA, implementando plantillas y prácticas comunes para una utilización coherente, predecible y de mayor calidad.



Piloto Automático

Las IA se encargan de tareas completas de, incluyendo pruebas y correcciones, con los profesionales actuando como supervisores del proceso.

¡Gracias!

diverger