UAI

Universidad Abierta Interamericana

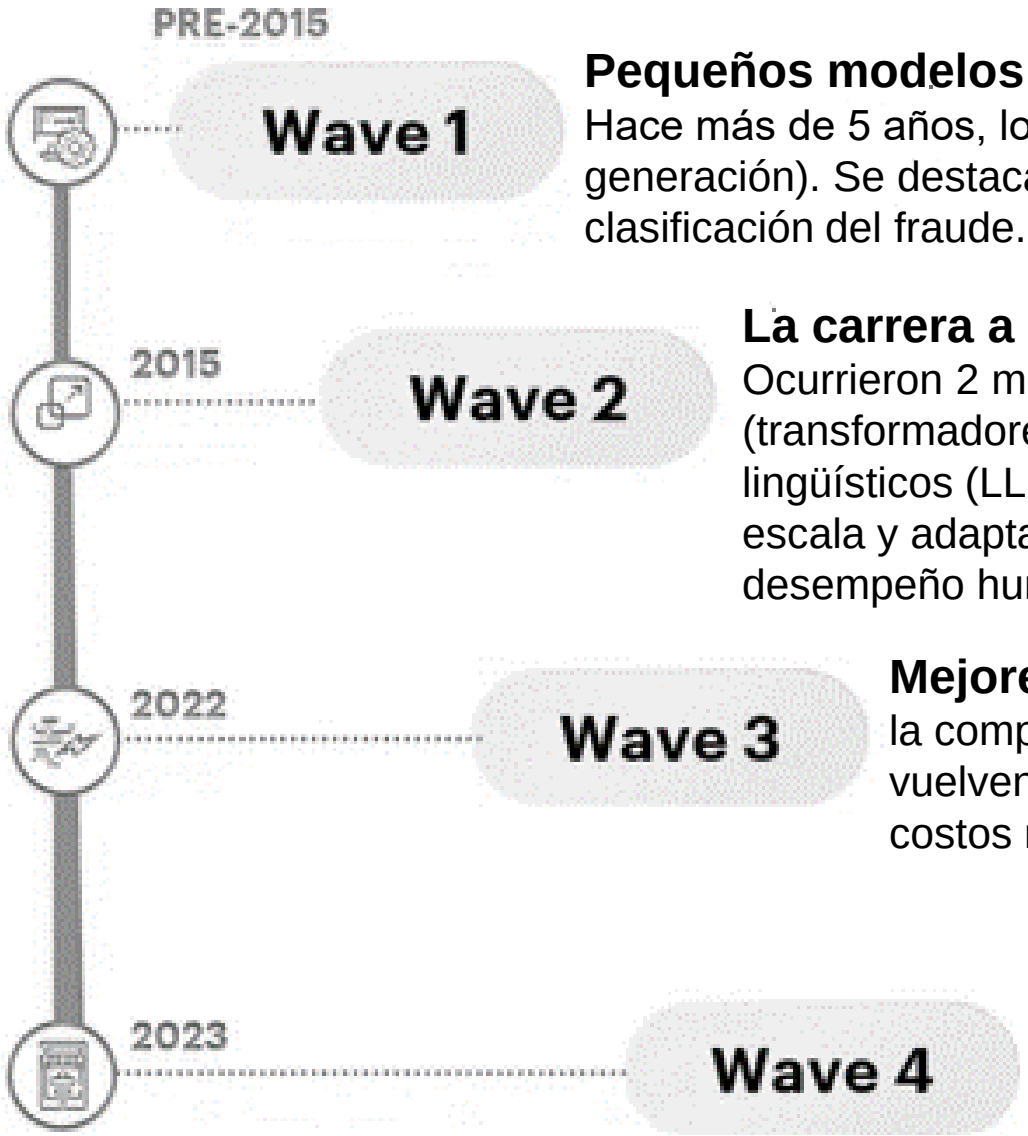
IA generativa en la Industria

CITII 24

Ing. Maximiliano Bonaccorsi

LLMs (large language models)

Generación de modelos: Better models, more data, more computer



Pequeños modelos reign supreme (Pre 2015)
Hace más de 5 años, los modelos pequeños se consideraban “State of the art” (de última generación). Se destacaron en tareas analíticas y permitieron predecir en tiempo real para la clasificación del fraude.

La carrera a escala (2015-Actual)
Ocurrieron 2 momentos históricos. Google introdujo el concepto de Transformers (transformadores) en PNL. Stanford clasificó acertadamente estos grandes modelos lingüísticos (LLms) como modelos fundamentales. Capacitados con datos amplios a escala y adaptables a una amplia gama de tareas posteriores. Los LLms imitaron el desempeño humano en el habla, la visión, la comprensión del lenguaje y la comprensión.

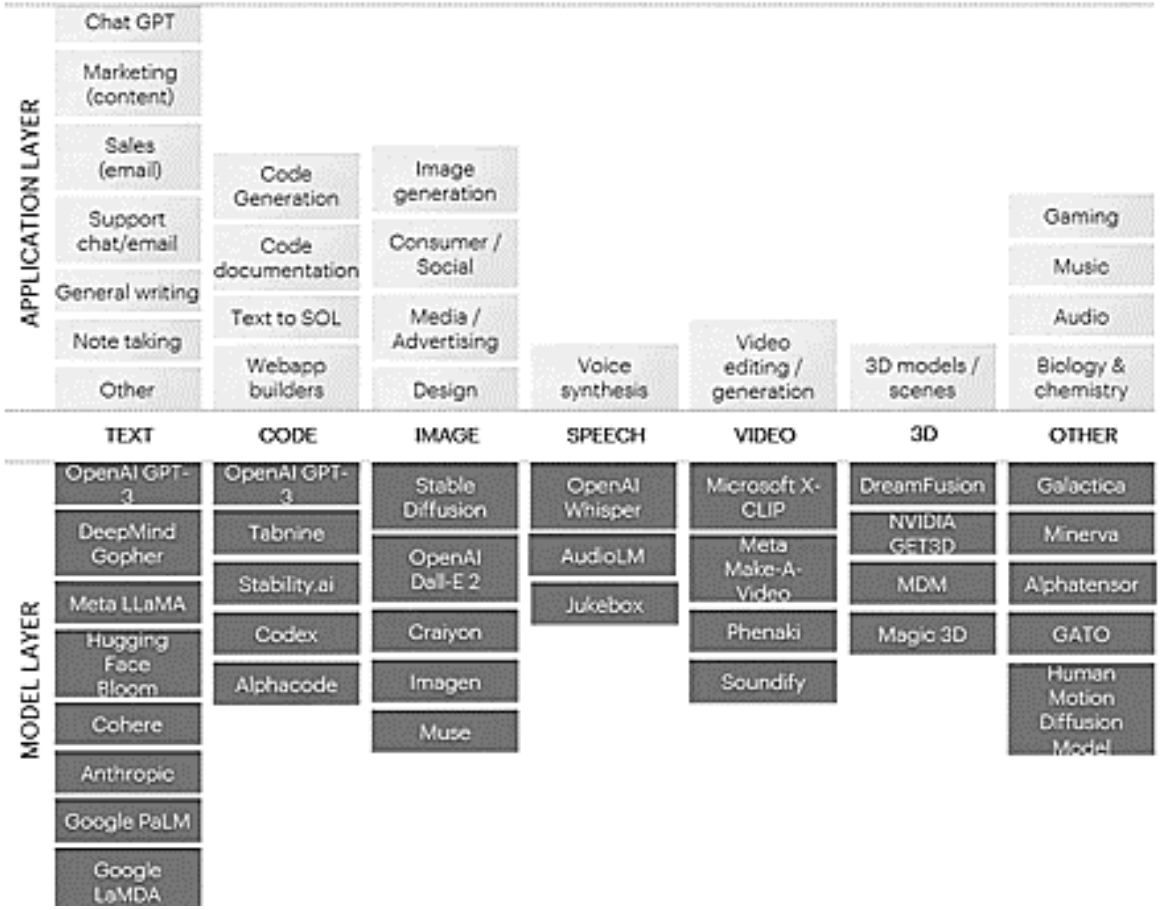
Mejores, más rápidos y baratos (2022+)
la computadora se vuelve más barata, los algoritmos y los modelos grandes se vuelven mejores. Nuevas técnicas, como los modelos de difusión, reducen los costos necesarios para entrenar y ejecutar inferencias.

Nuevas Aplicaciones revolucionarias (Actual)
Los LLms continúan volviéndose mejores, más rápidos y más baratos y Así los dispositivos móviles fomentaron un mercado nuevo aplicaciones innovadoras utilizando IA generativa.

LLMs (large language models)

Modelos y apps de IA generativa

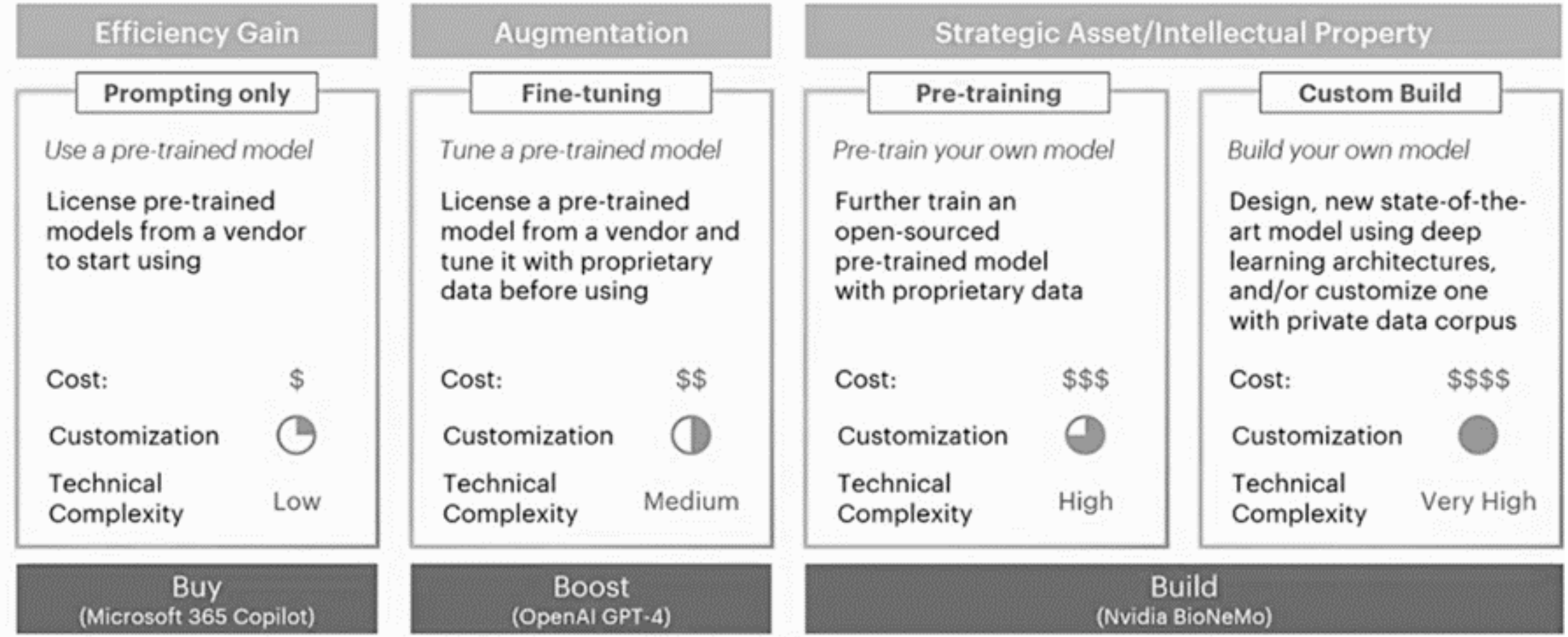
Tipo	Descripción
Modelos de texto	Decente en escritura corta/forma media. Necesidad de resultados de mayor calidad, contenido más extenso y un mejor ajuste específico del dominio
Generación Código	Gran impacto sobre la producción del desarrollo (ej. Github Copilot)
Imágenes	Advenimiento de diferentes estilos estéticos y técnicas de edición y modificación.
Speech synthesis	por un tiempo (¡hola siri!), pero las aplicaciones empresariales y de consumo están mejorando
Video y 3d models	Potencial para destrabar mercados creativos como cine, juegos, VR, arquitectura y diseño de productos físicos
Otros Dominios	Modelos Fundamentales (LLMs) ocurren a través de muchos campos desde la música hasta las farmacéuticas



IA generativa

Enfoques para habilitar modelos de base (foundation models)

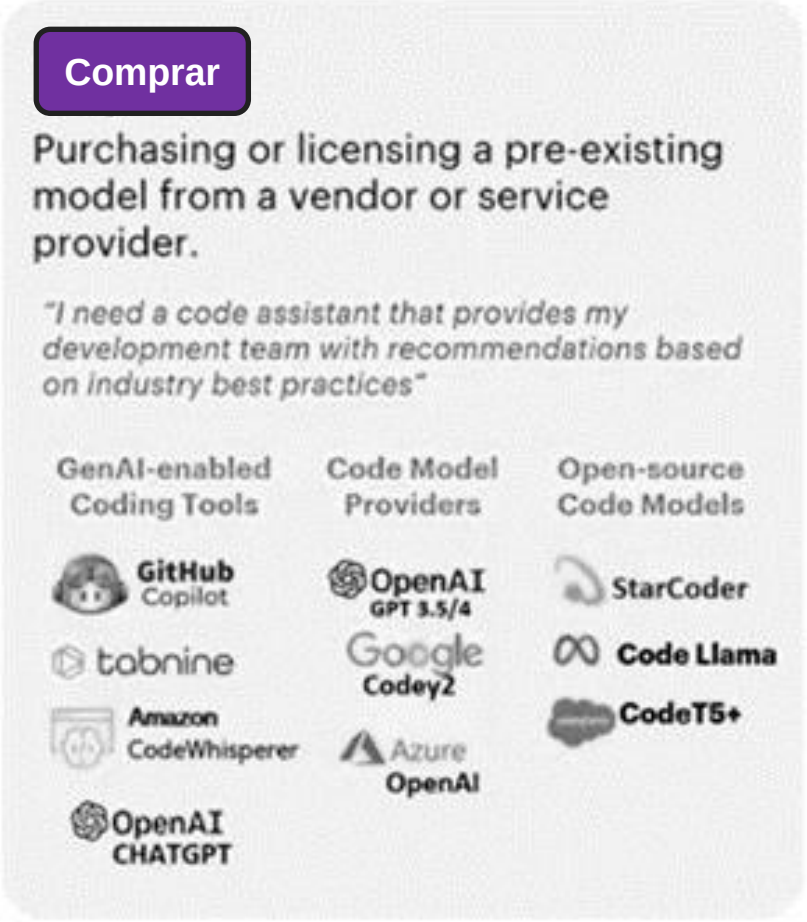
Hay 3 enfoques principales para habilitar modelos de base y su complejidad técnica y costos aumentan con el nivel de personalización requerido.

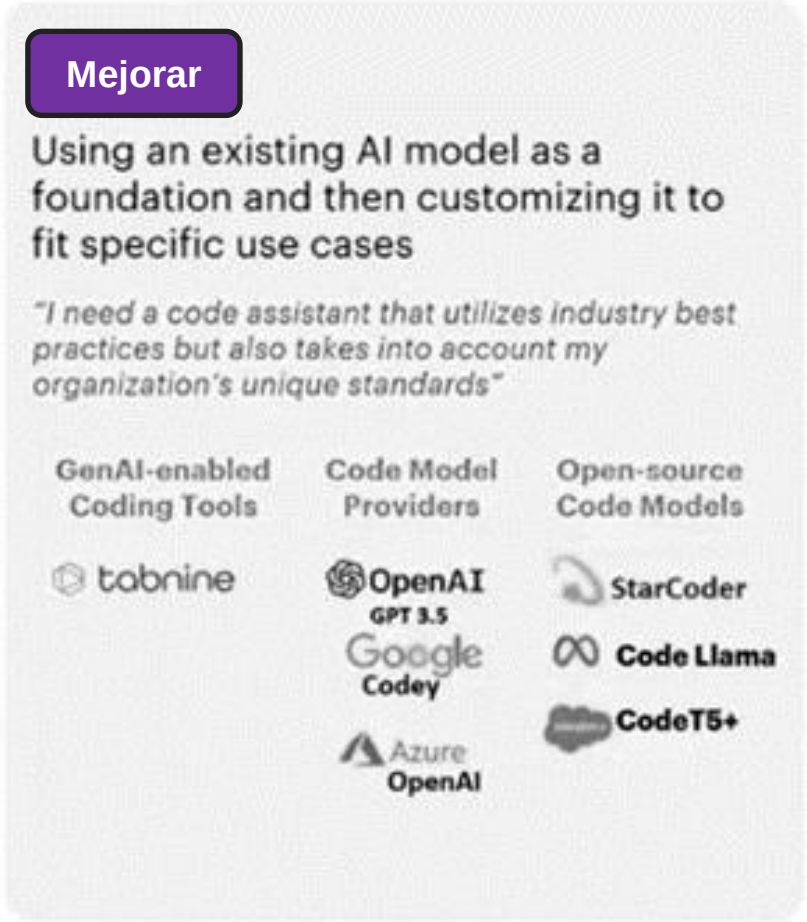


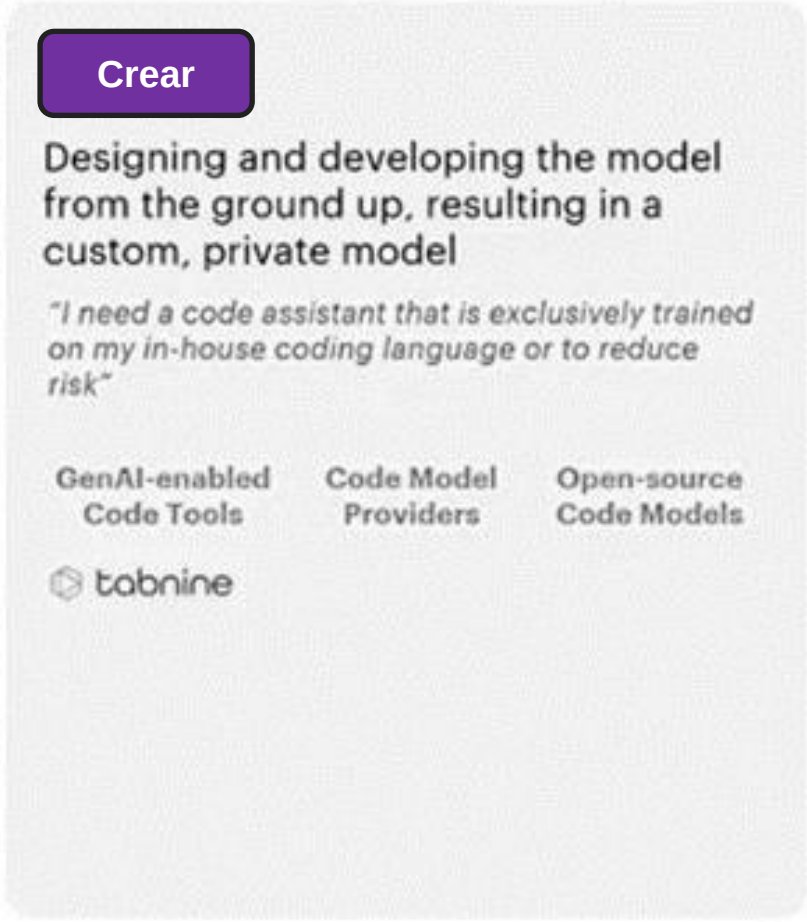
IA generativa

Ejemplo: Enfoques de habilitación de modelos para codificación asistida por IA

Se debe encontrar la combinación correcta de enfoques para impulsar el valor empresarial y crear su propia llm para tareas específicas







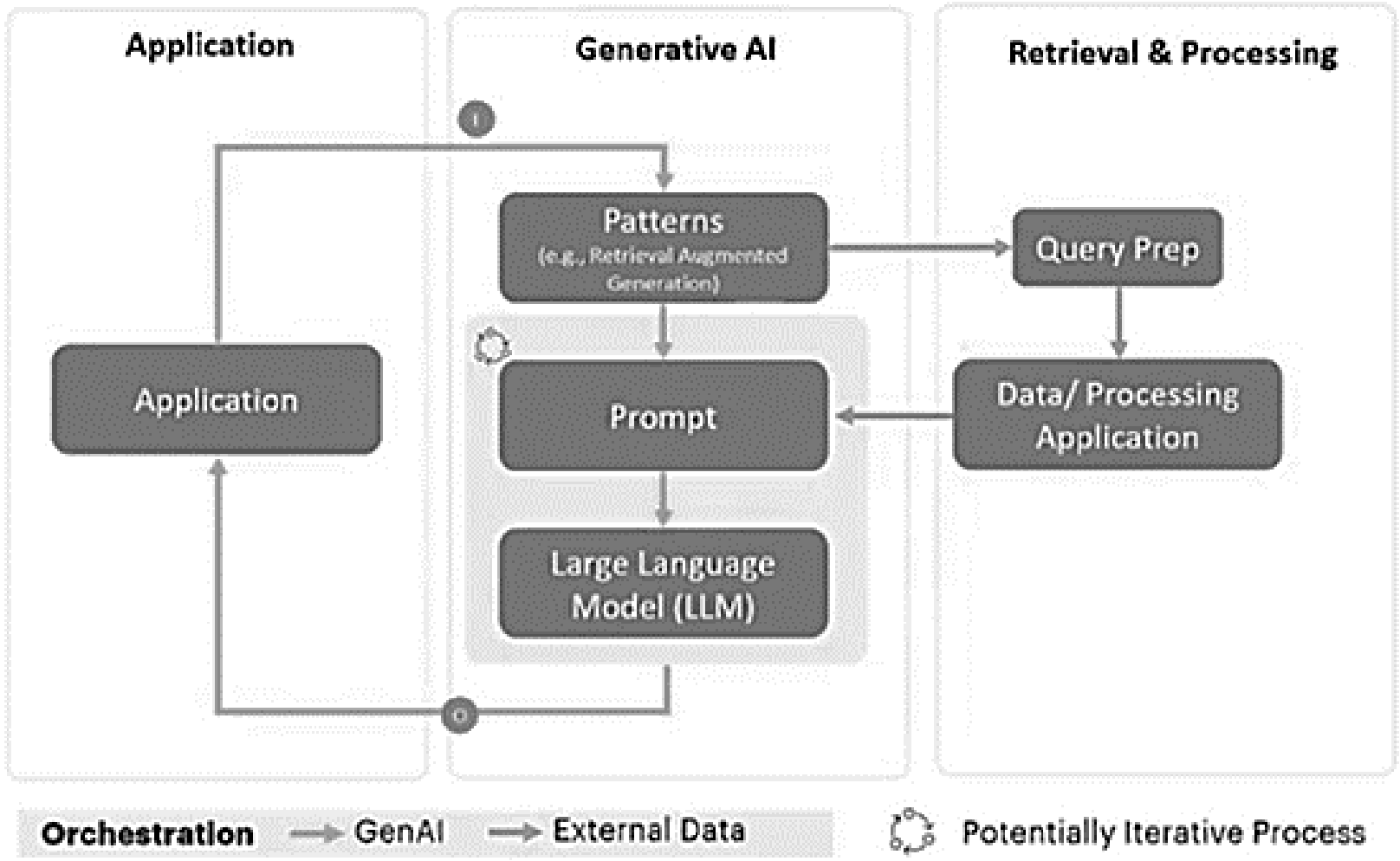
IA generativa

Mejorar

La siguiente fase agrega datos para habilitar la asistencia por copilot.

Ejemplo: patrón de generación aumentada de recuperación aplicado a la entrega de tecnología.

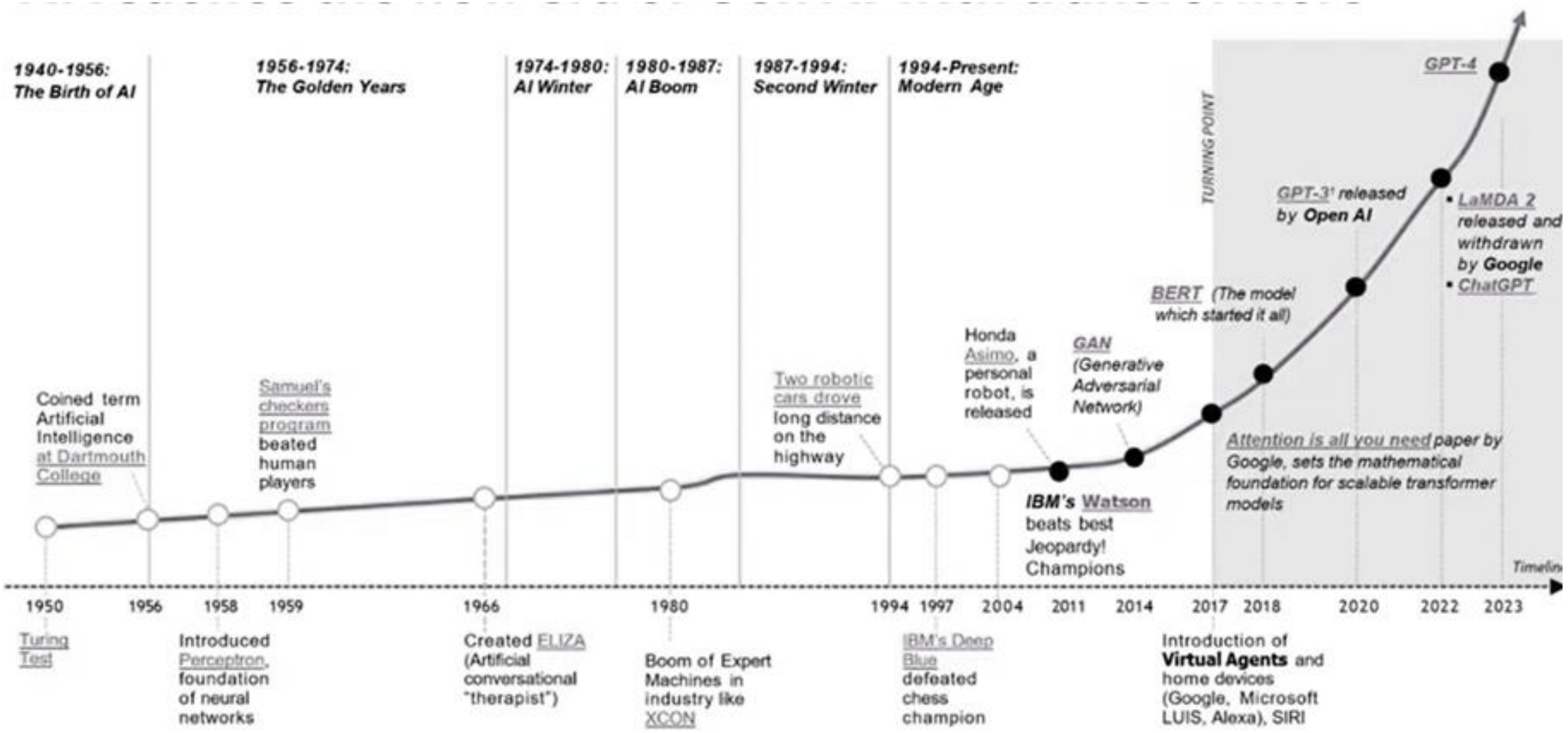
Outputs
- Code
- Documentation
- Tests
- Ticket Resolution
- Requirements
- Workflows
- Scripts
- Pipelines
- Q/A driven insights



Knowledge Base
- Software Code & Configuration Files
- Infrastructure as Code Templates
- Architecture
- Design Specification
- Vendor manuals
- Requirements
- Workflows
- System Logs
- Standard Operating Procedures
- Business Process

IA generativa

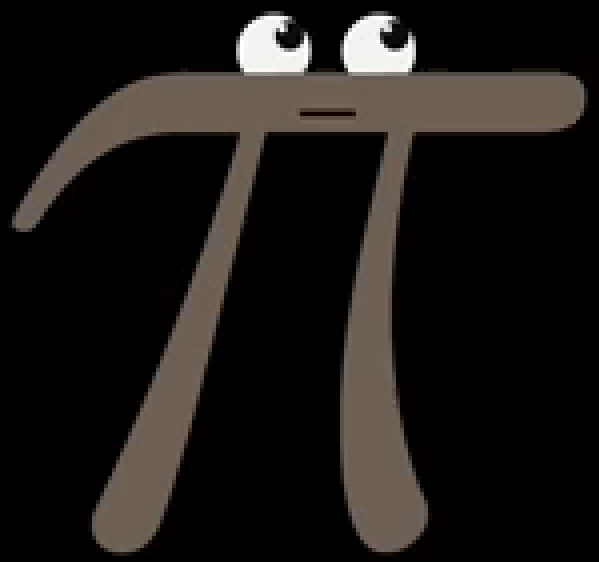
AI llega a la nueva era de gen ai con transformers



IA generativa

la nueva era de IA Generativa surge con Transformers

Veamos su funcionamiento

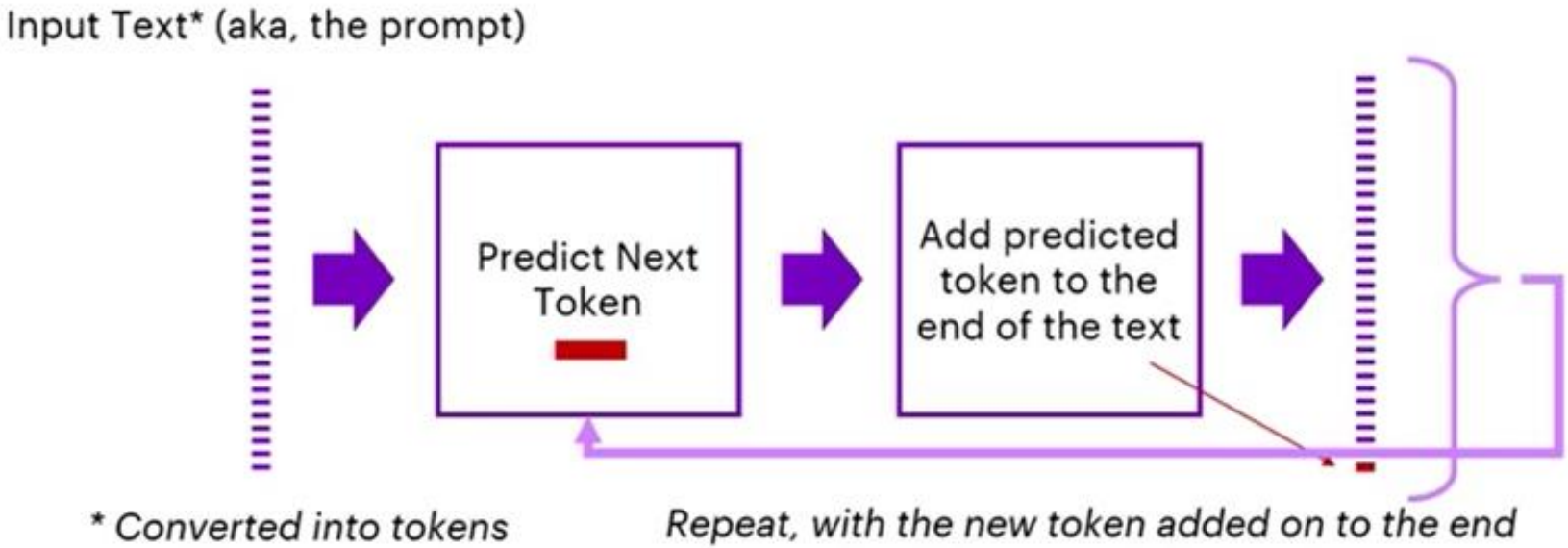


IA generativa

¿Como funciona Gen AI?

Parte 1- Predicción

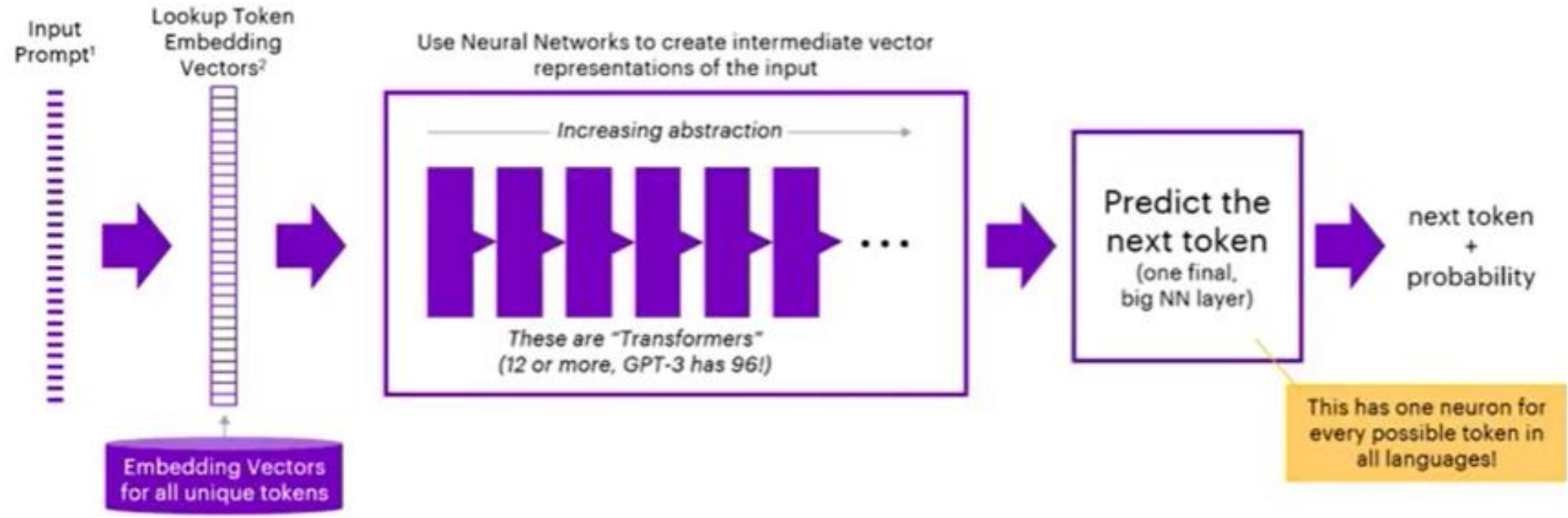
Ejemplo de Token:



IA generativa

¿Como funciona Gen AI?

Parte2- Transformers



Nota:

1. Convertido a Token
2. Cada token único tiene un vector, cada vector es una lista de 700 a 13000 números de punto flotante(el valor del vector actual para cada token único esta entrenado por el proceso de entrenamiento NN.
3. El vector del token se modifica según la posición dentro del prompt (mensaje).

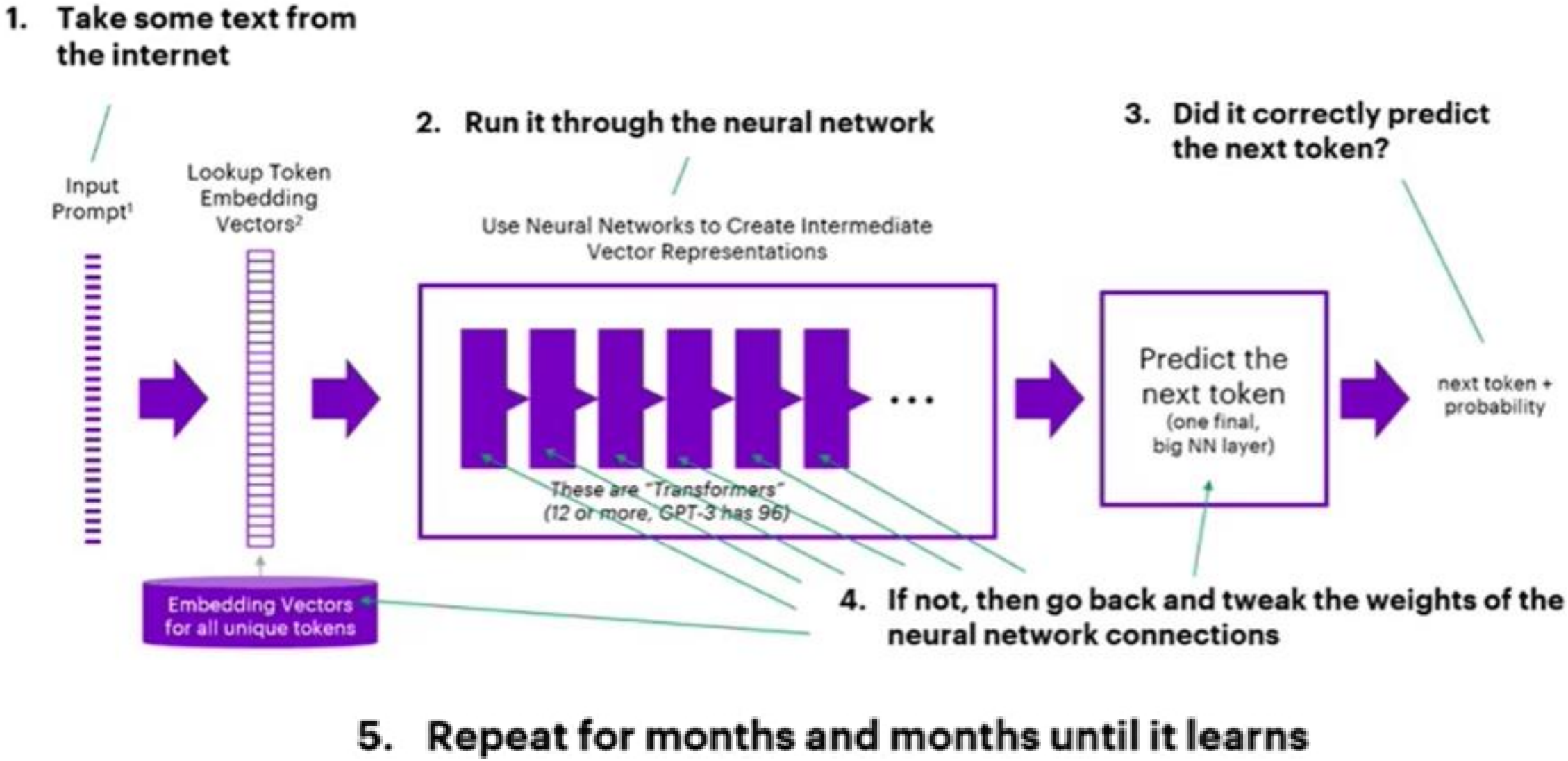
Cada Transformer:

1. combina todos los vectores simbólicos en el input (entrada) entre sí (esta es llamada la “self attention layer” y este complejo mashup permite que la red neuronal aprenda la sintaxis)
2. tiene múltiples (al menos 2) capas de red neuronal en su interior (con millones de neuronas y billones de pesos de conexiones de red neuronal interconectadas

IA generativa

¿Como funciona Gen AI?

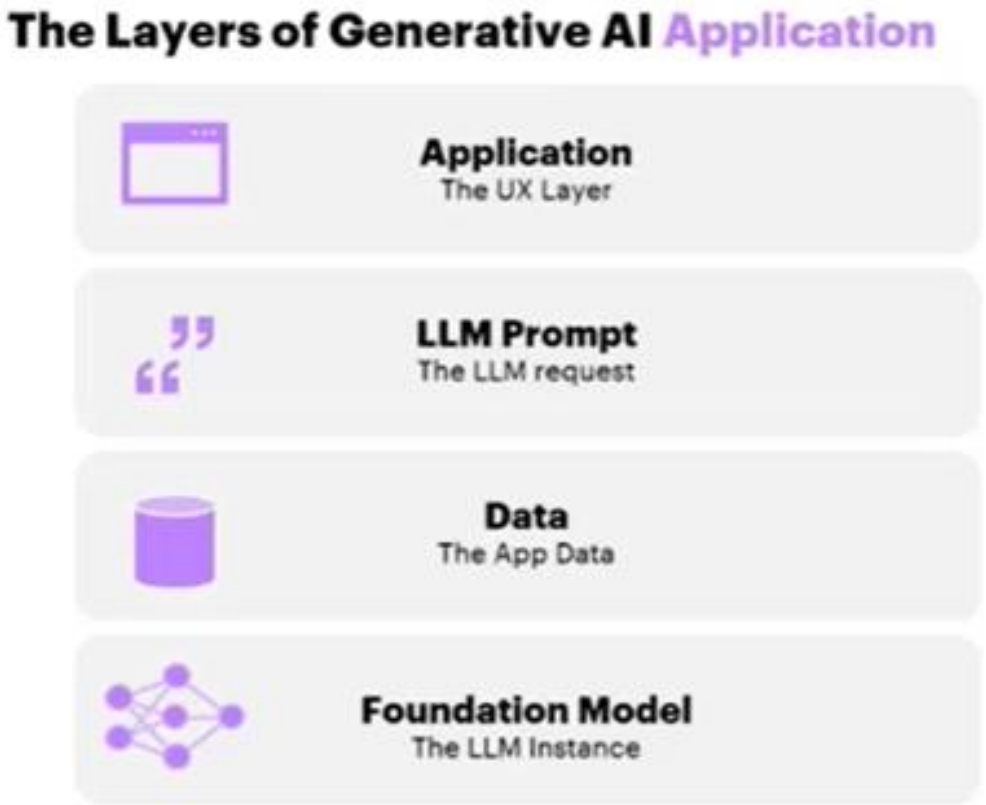
Parte3- Training

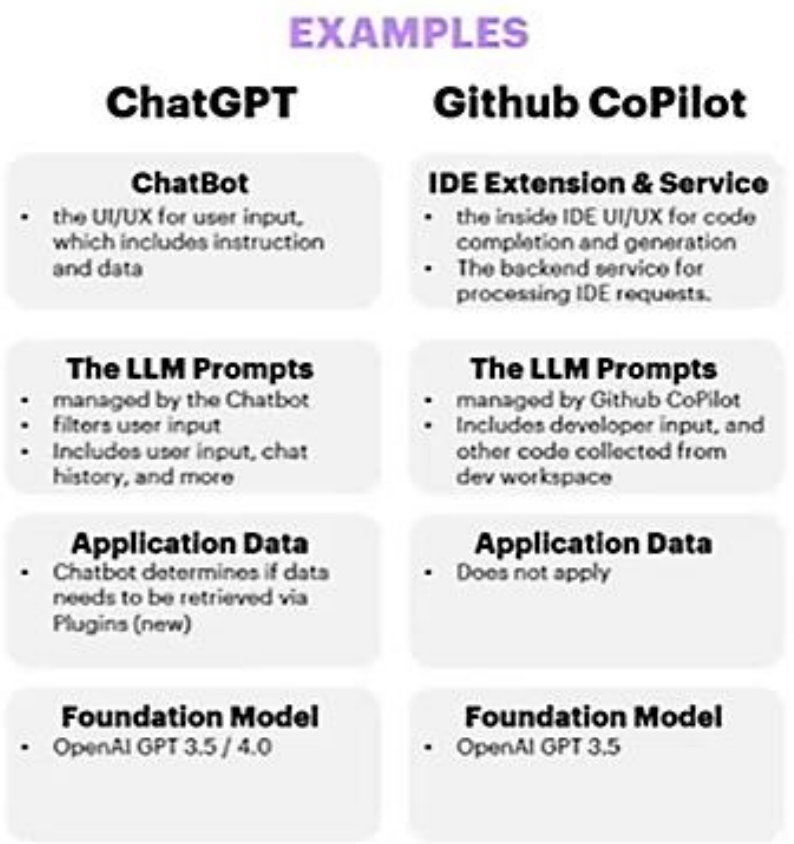


IA generativa

Utilización de modelos MMLs en aplicaciones

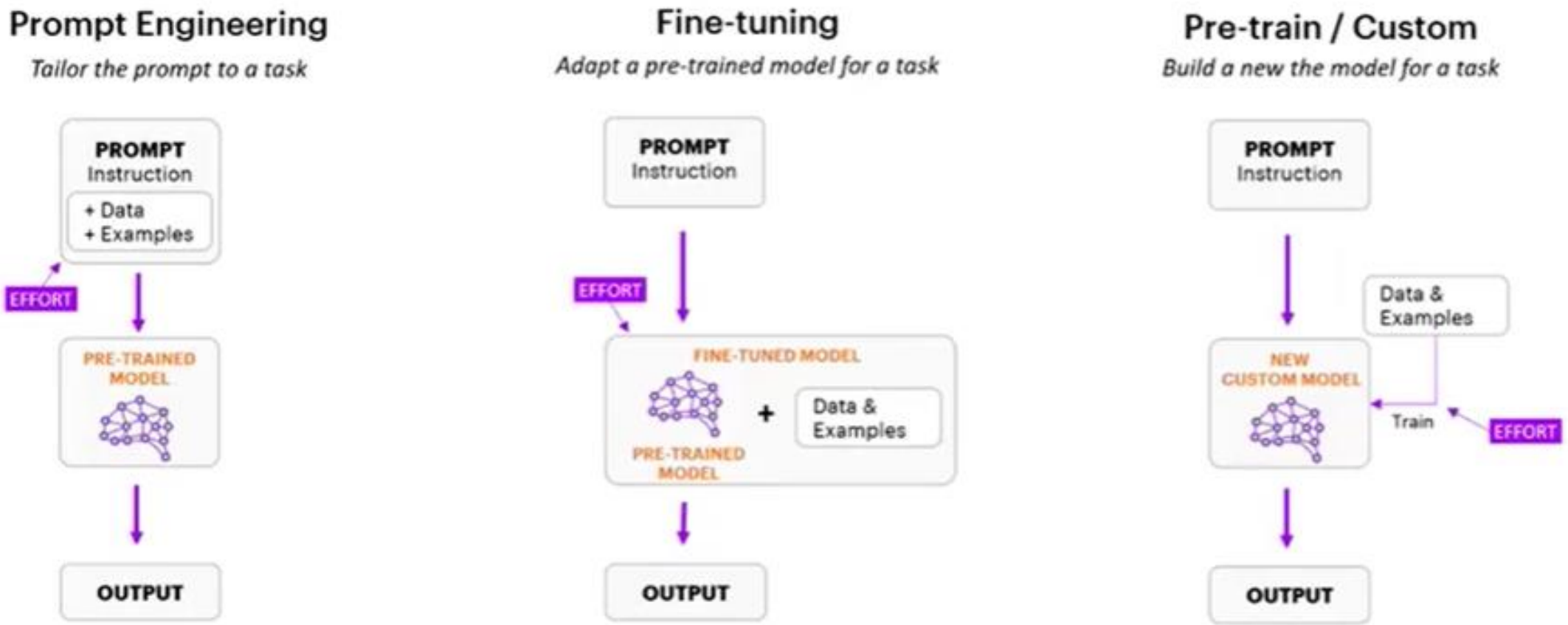
El Flujo operacional para una aplicación no difiere en
Diferentes tipos de aplicaciones





IA generativa

3 enfoques clave para adaptar los LLms a necesidades específicas



La complejidad técnica y los costos escalan hacia arriba según el nivel de customización que sea requerida

IA generativa

3 enfoques clave para adaptar los LLms a necesidades específicas

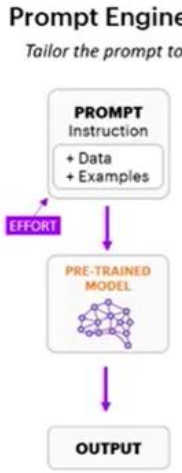
adaptar el mensaje a una pregunta

La complejidad técnica y los costos escalan hacia arriba según el nivel de customización que sea requerida

IA generativa

Prompt engineering es la forma de optimizar los resultados Gen IA

Promp engineering es una técnica utilizada en LLms para guiar o dirigir su producción proporcionándoles un mensaje específico, un conjunto de instrucciones o datos, donde el objetivo es mejorar la precisión y relevancia de la salida del modelo.



01. Content

Give it content to process

Often, the machine will need information to perform a task. Include this in the prompt.

- *Document text*
  - *E.g., Documentation, policies, procedures, etc.*
- *Tabular data (e.g., CSV)*
- *Lists of things to choose from (e.g., APIs or data sets)*

02. Instructions

Tell it what to do

This is where you tell it what you want to accomplish.

- *Answer question*
- *Create a plan*
- *Summarize some content*
- *Process some content*
- *And, most importantly, what sort of output you want*

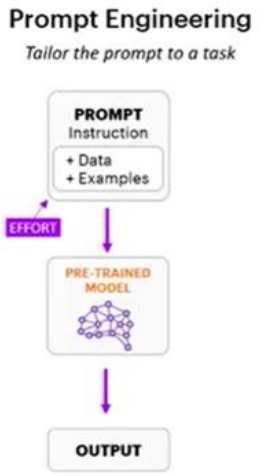
03. Examples

Provide examples, if needed

If it is unable to follow instructions, you may need to provide examples.

- *Content understanding*
- *Aspects or dimensions of interest*
- *What to extract*
- *Output formatting*

IA generativa



Hay 3 técnicas de aprendizaje en contexto.

Entender las técnicas de aprendizaje en contexto provee instrucciones adicionales para ayudar la interpretación

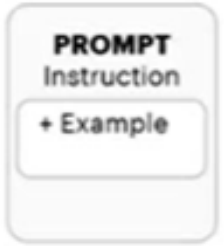
Técnicas de aprendizaje en contexto: definiendo los datos limitados.

Zero shot learning

Zero-shot learning aims to correctly interpret tasks without any specific examples of those tasks during training, purely relying on the ability to understand instructions and context.

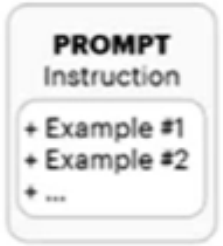
One-shot learning

One-shot learning focuses on the ability of a model to accurately make predictions after seeing just a single training example.

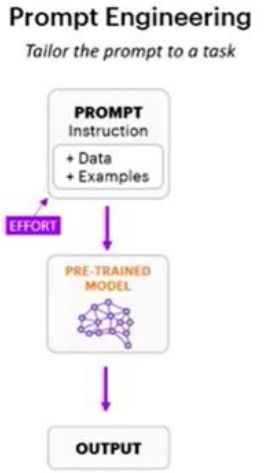


Few shot learning

Few-shot learning allows a model to generalize or make predictions after seeing a few examples.



IA generativa

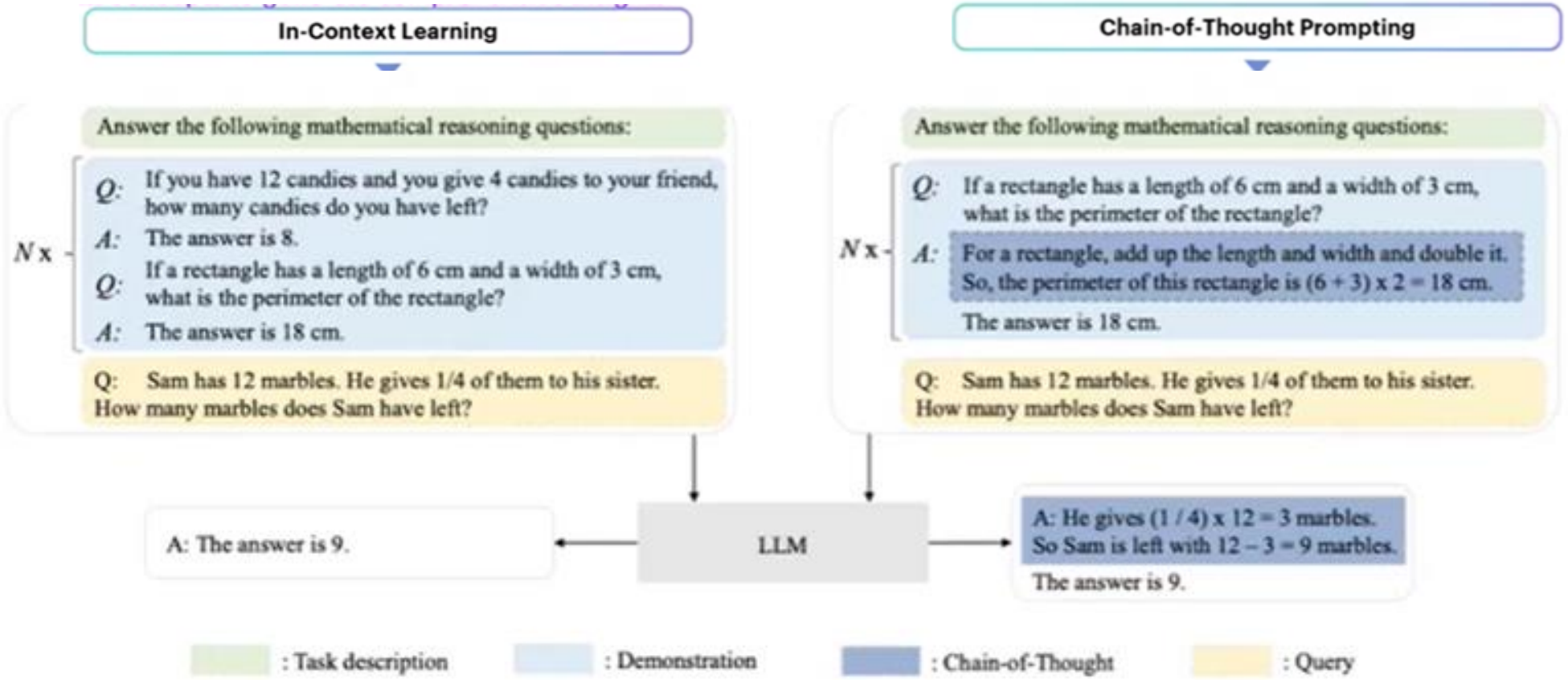


Adaptando LLms con prompt engineering.

Un modelo aprende a través de ejemplos analizando los datos de entrada (inputs), identificando patrones y realizando predicciones o decisiones, mientras que el razonamiento en cadena de pensamiento permite que el modelo se base en conocimientos previos, conectando diferentes conceptos para generar conocimientos integrales.

aprendizaje en contexto

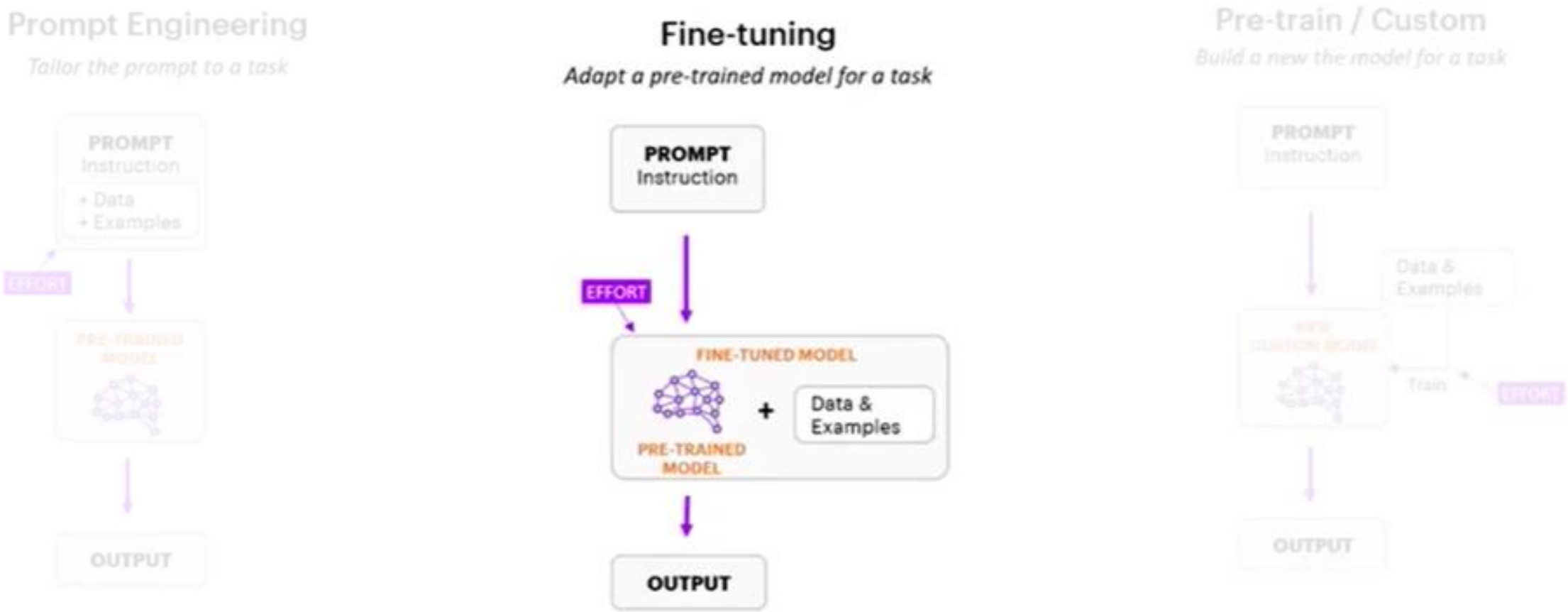
cadena de pensamiento



IA generativa

3 enfoques clave para adaptar los LLms a necesidades específicas

adaptar el modelo pre-entrenado para una tarea

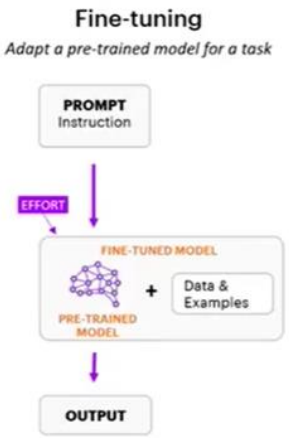


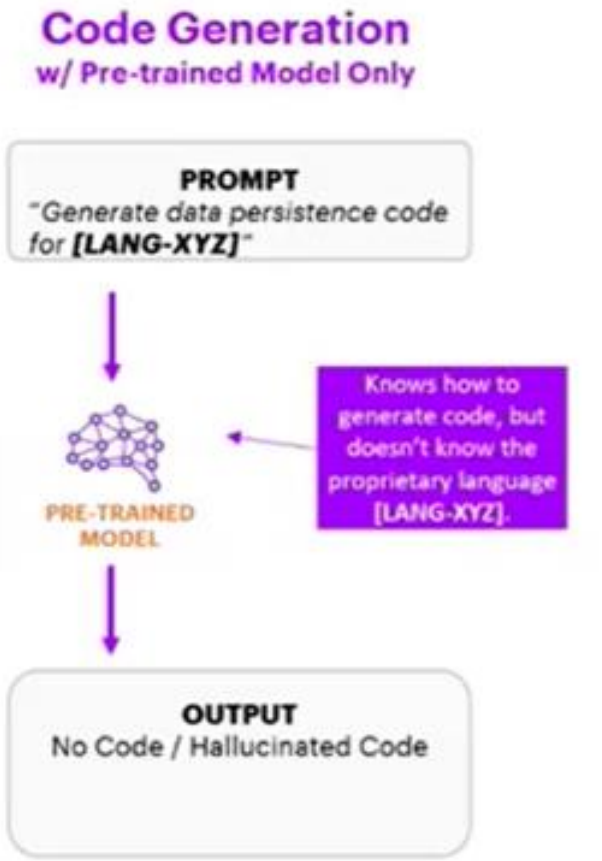
La complejidad técnica y los costos escalan hacia arriba según el nivel de customización que sea requerida

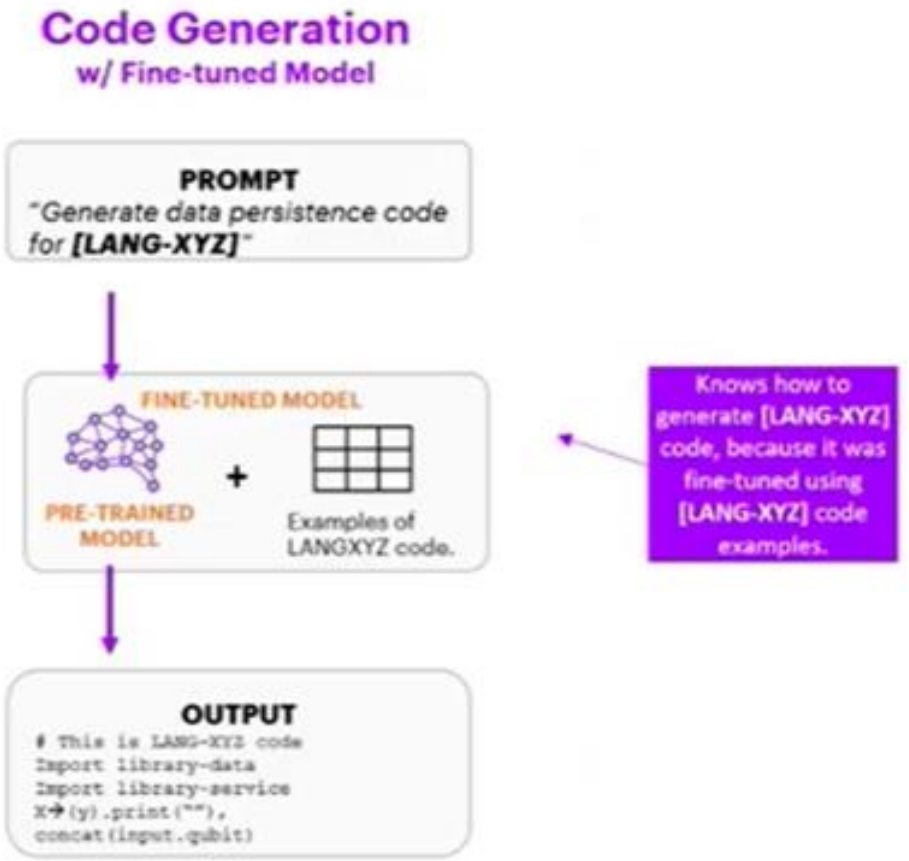
IA generativa

Adaptar LLms a necesidades específicas con Fine-tune

Las habilidades de los modelos se pueden adaptar aún más (o fine-tune) para adaptarse a casos de uso únicos después del entrenamiento.



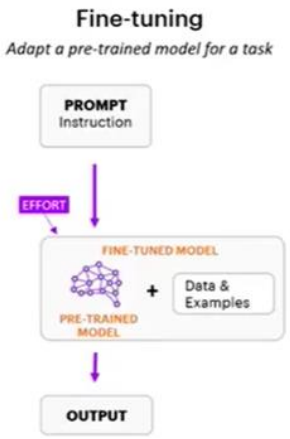


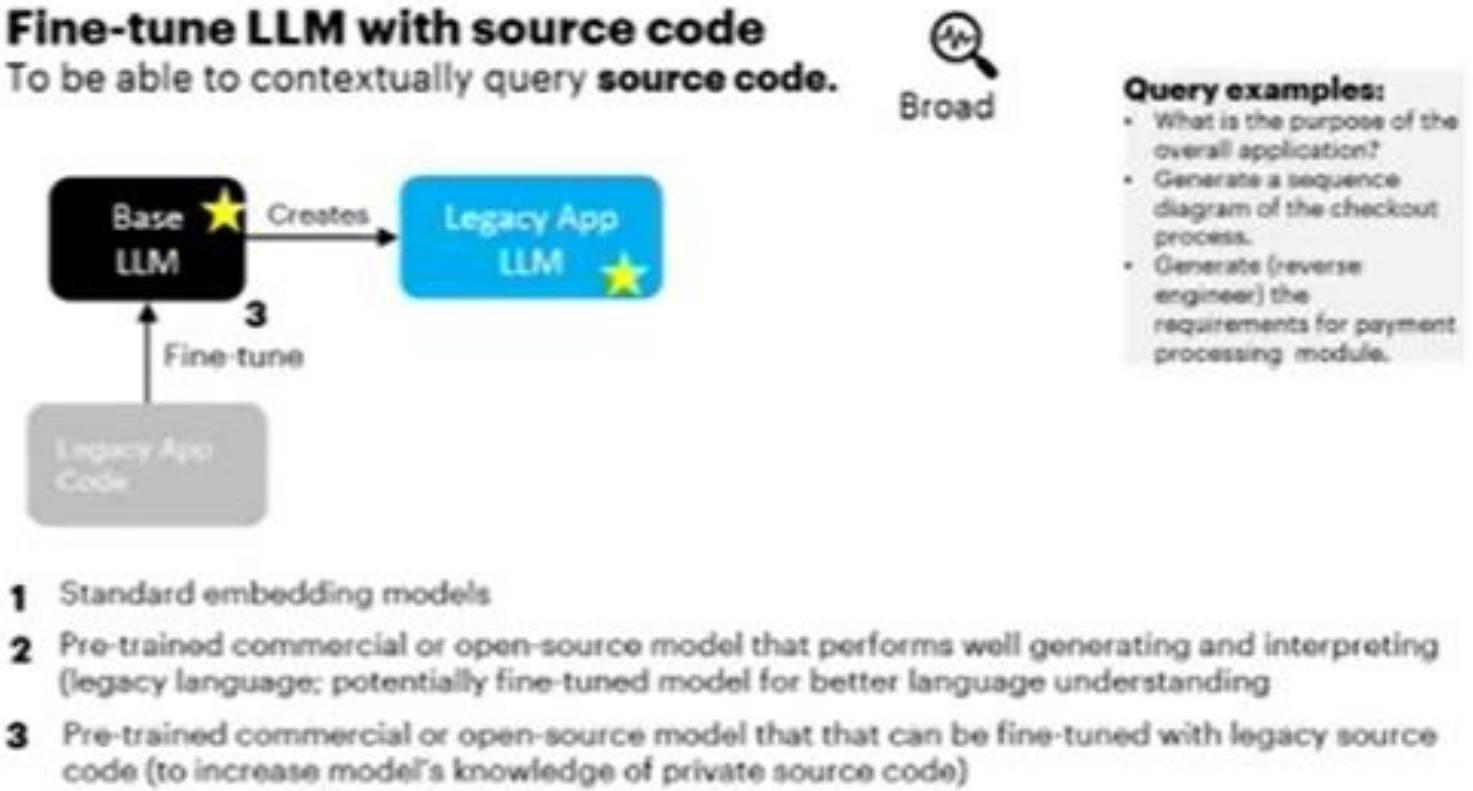


IA generativa

Adaptar LLms a necesidades específicas con Fine-tune

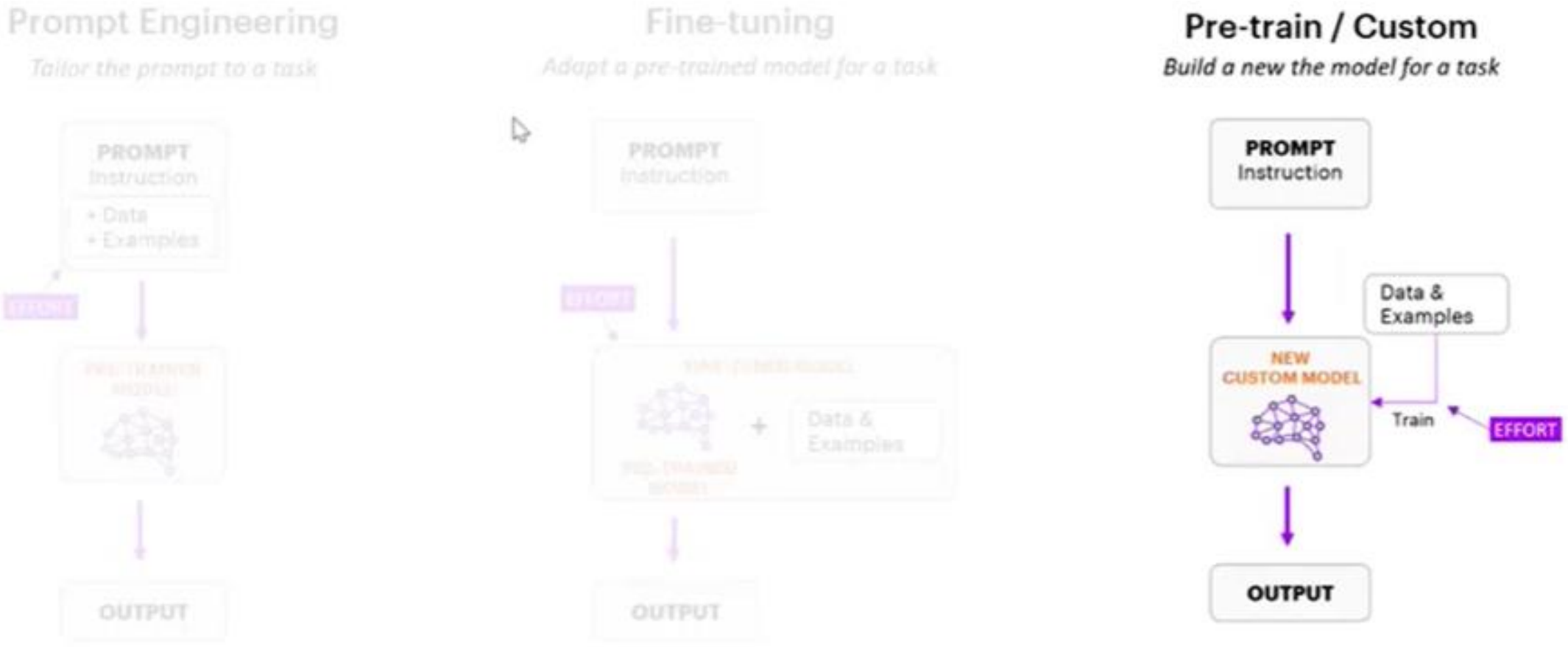
Las habilidades de los modelos se pueden adaptar aún más (o fine-tune) para adaptarse a casos de uso únicos después del entrenamiento.





IA generativa

3 enfoques clave para adaptar los LLms a necesidades específicas



La complejidad técnica y los costos escalan hacia arriba según el nivel de customización que sea requerida

IA generativa

Componentes arquitectónicos clave de la aplicación Gen AI.

Una aplicación gen IA es una implementación de uno o más patrones de uso gen ia. Prompt engineering se utiliza para construir la indicación con las instrucciones, datos y ejemplos necesarios para que LLms complete el objetivo o la tarea.

Application
- The interface for end-users to access features enabled by GenAI

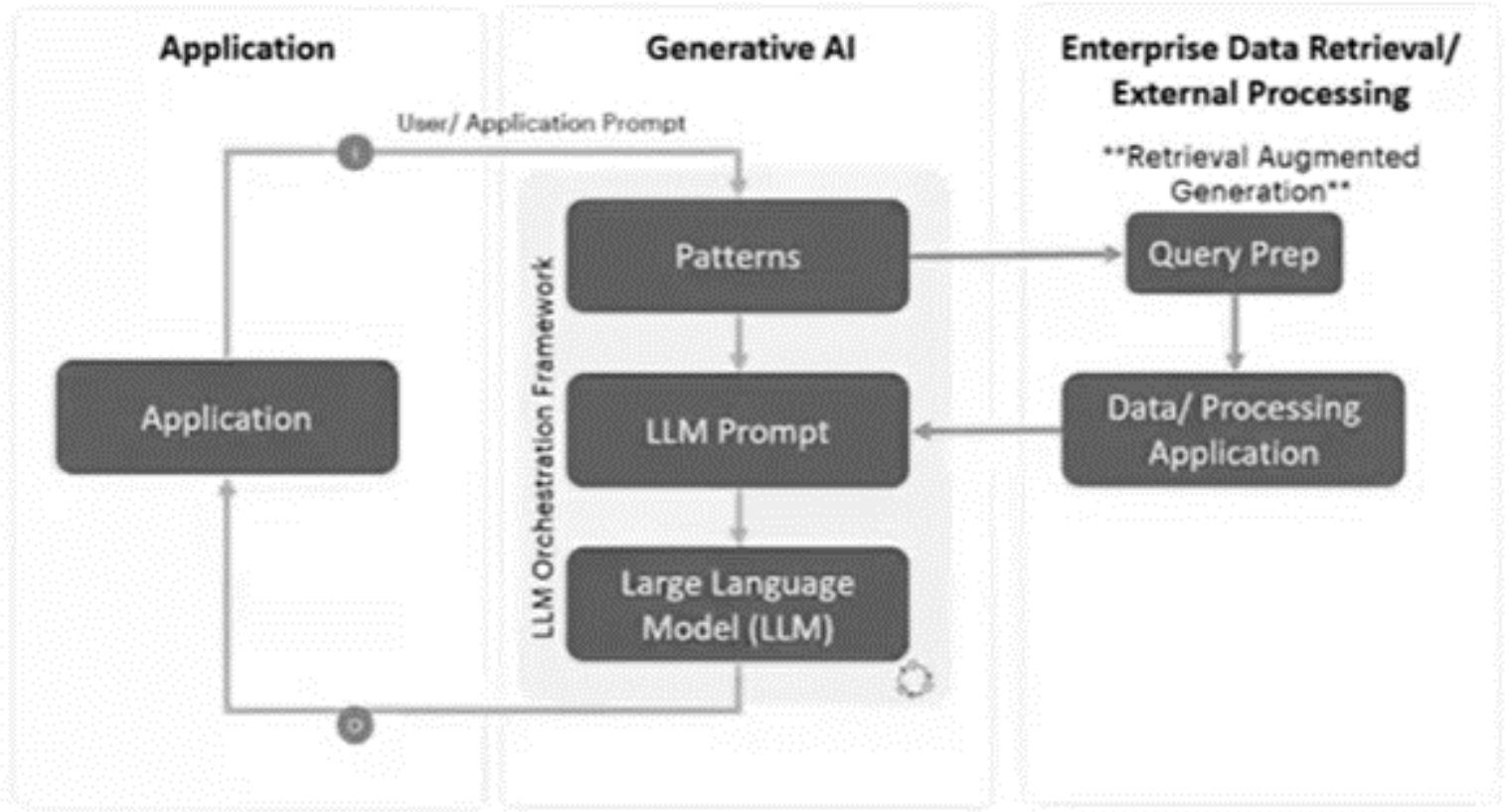
Patterns
- The generative AI pattern applied.

Query Prep
- The preparation of query to enterprise data store.

Data
- The querying the enterprise application or data store.

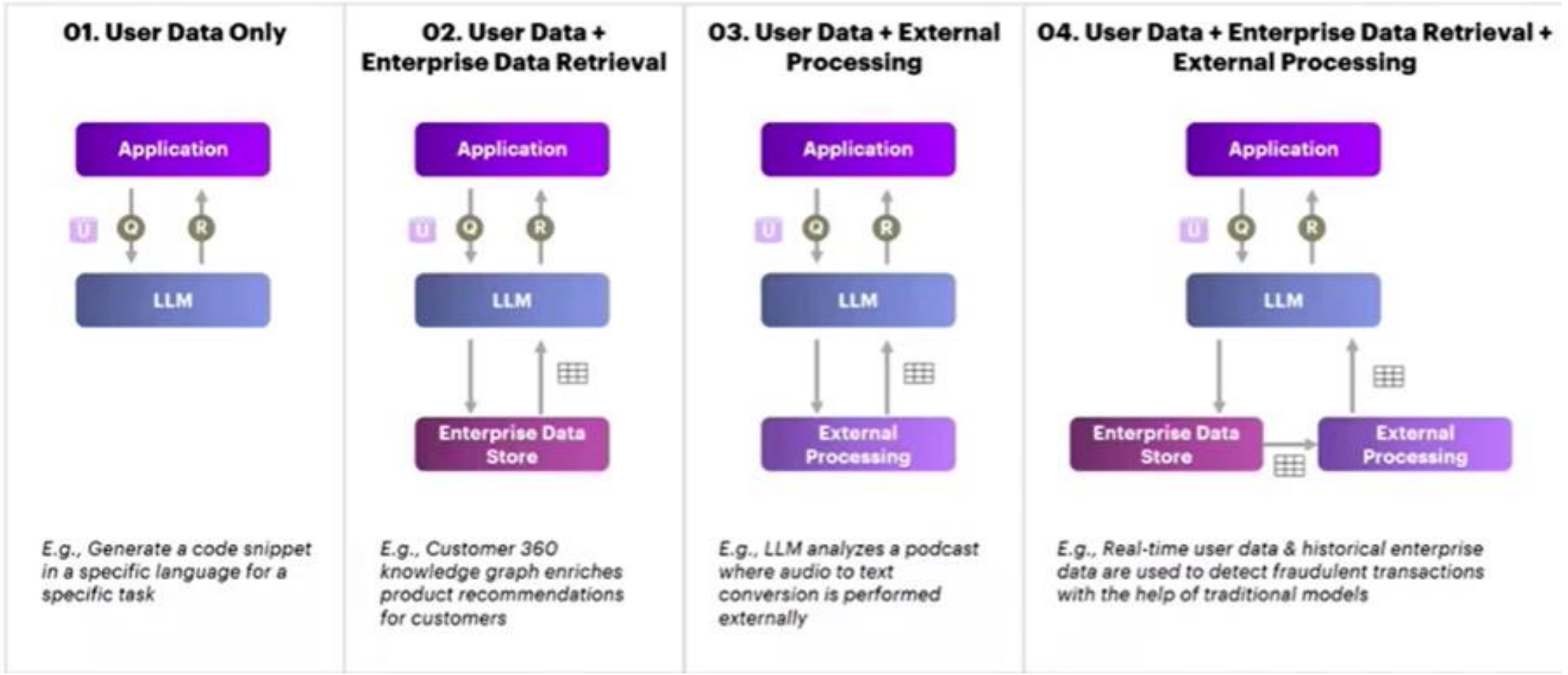
Prompt
- What is sent to the LLM
- An instantiation of a pattern
- includes original query, the history, the retrieved data.

Large Language Model (LLM)
- Generates (infers) an answer to question from data embedded in the LLM
- The LLM could be universal model, fine-tuned, or custom



IA generativa

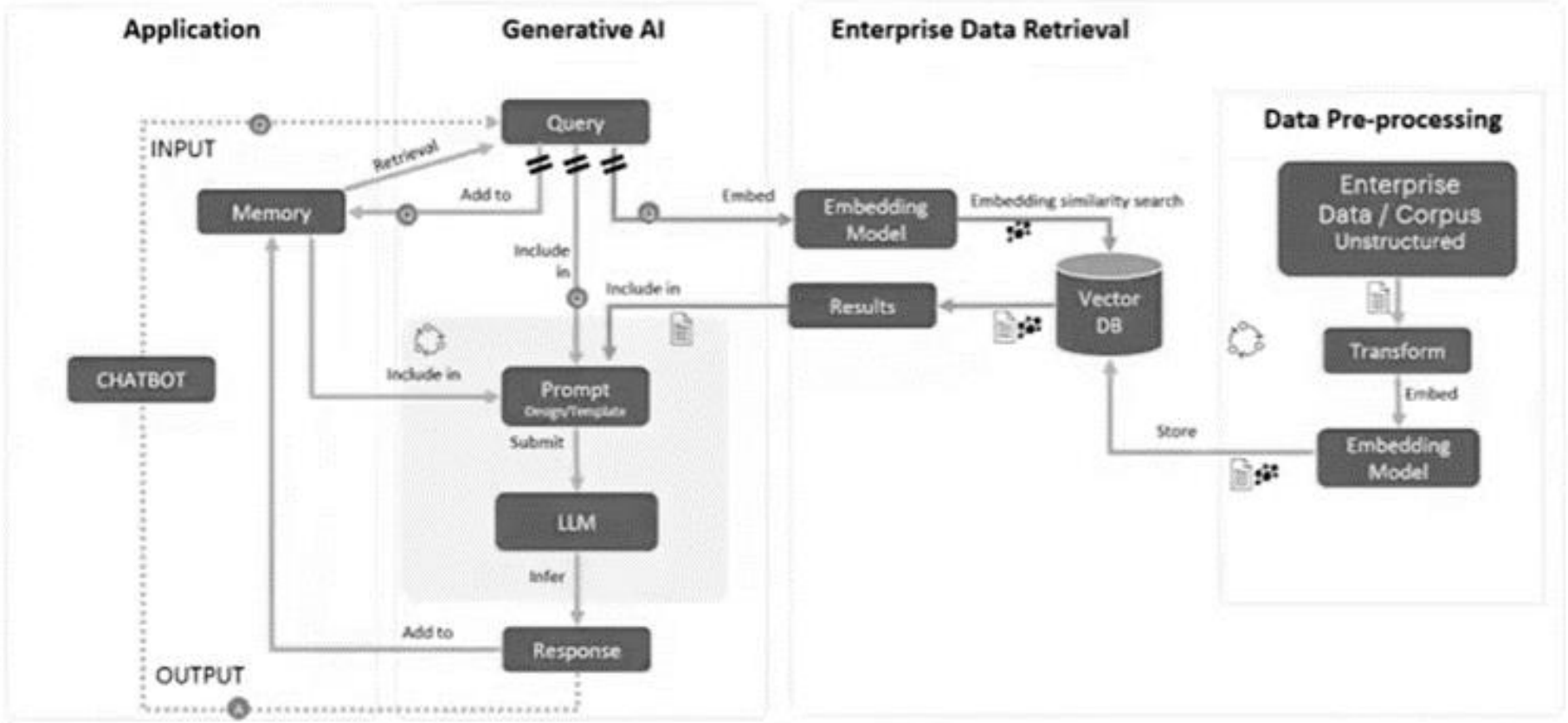
4 arquetipos centrales para la Gen AI

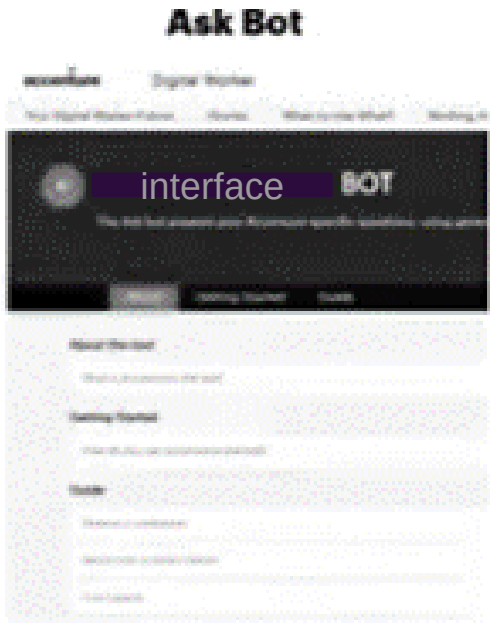


IA generativa

Búsqueda Semántica / Vector de búsqueda

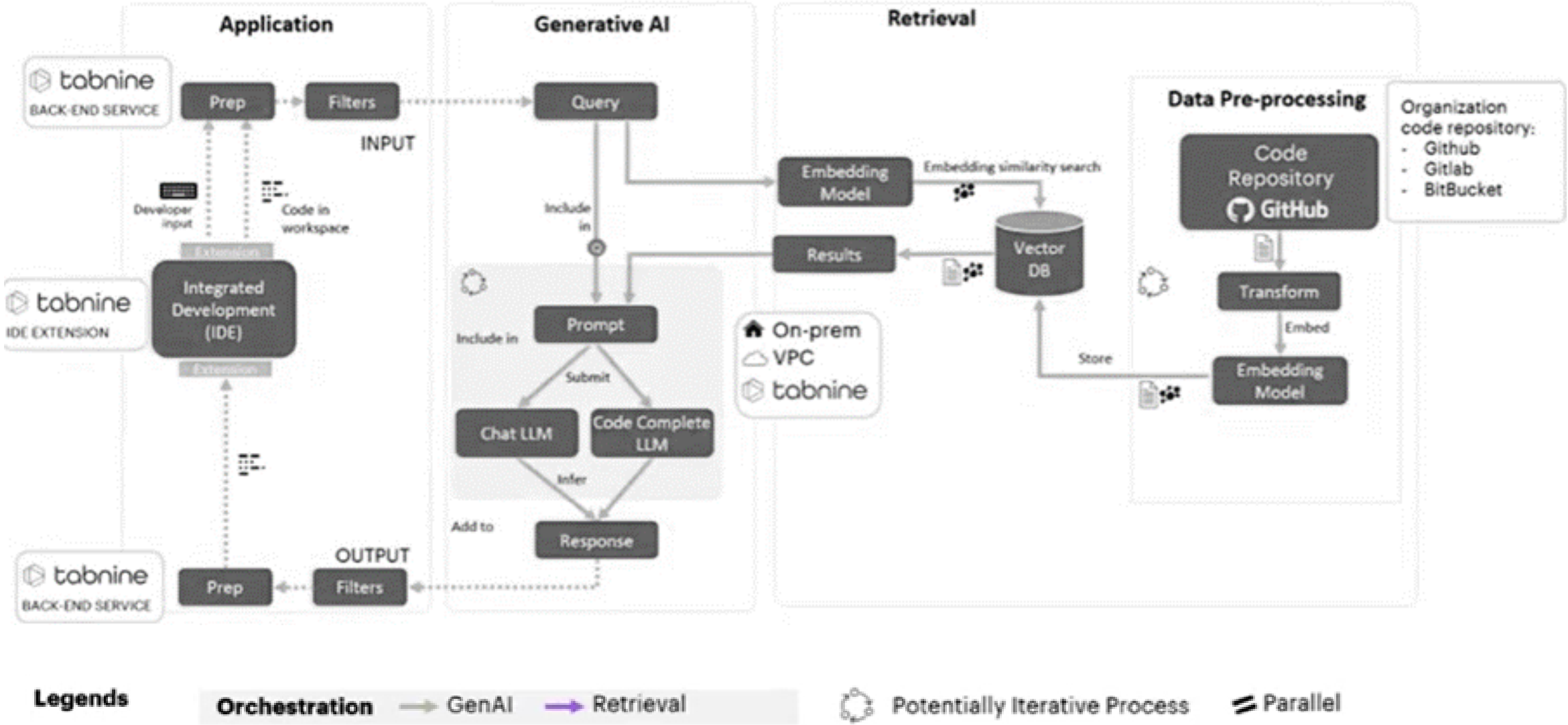
La recuperación de datos empresariales incorpora la consulta para buscar semánticamente en la base de datos vectorial almacenes de artículos que contienen información relevante, que se agregará como contexto al mensaje.





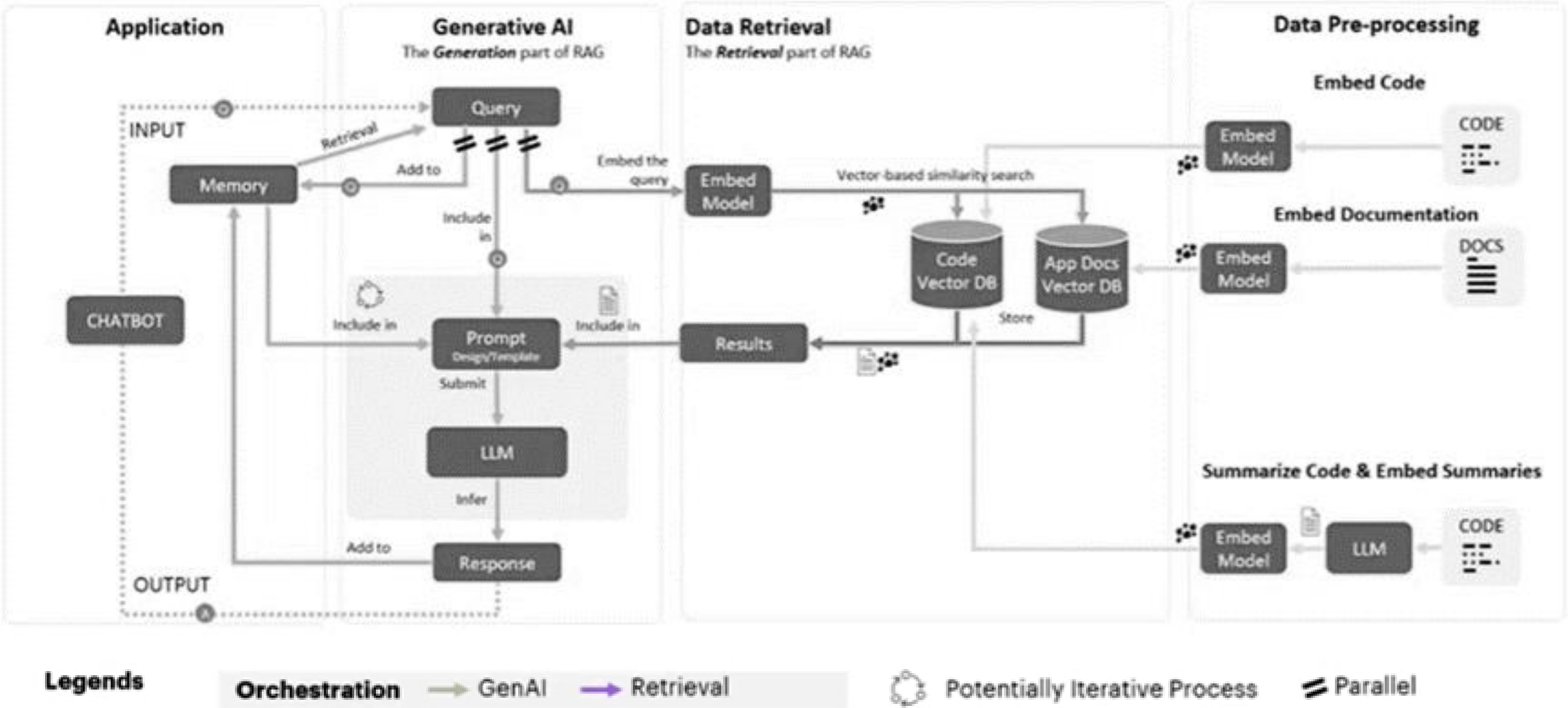
IA generativa

Ejemplo: Asistente de código (generación y completar código)



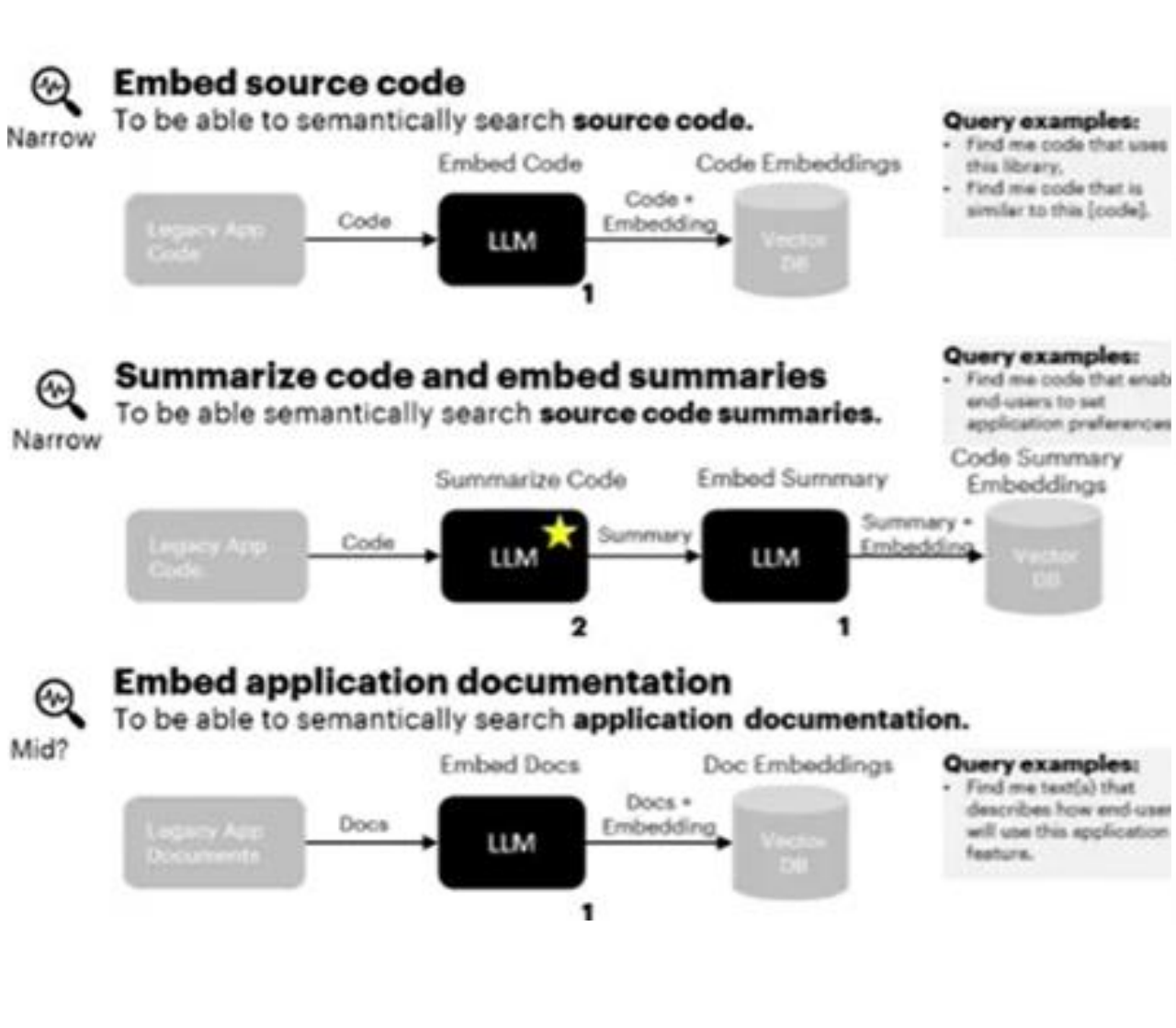
IA generativa

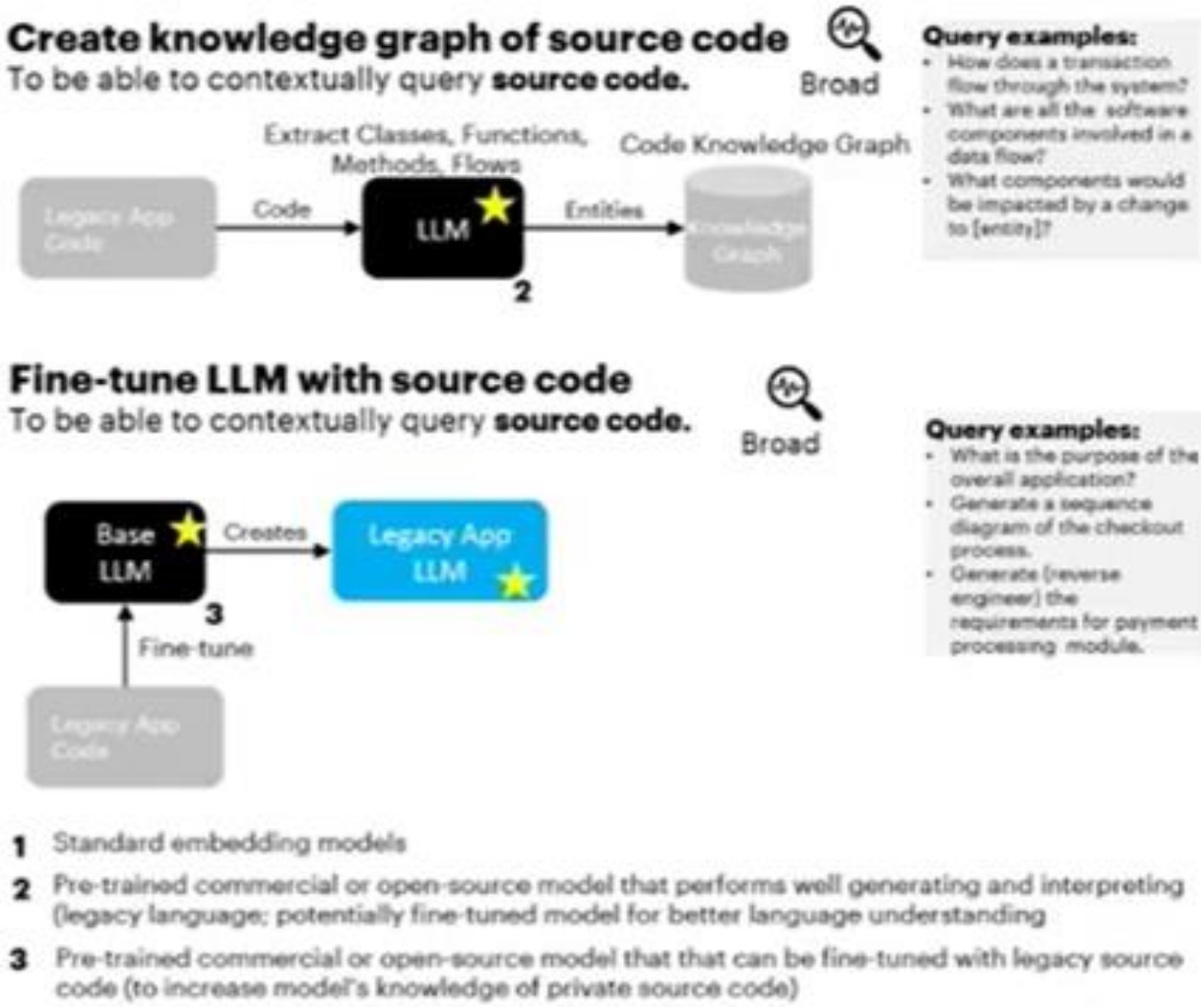
Ejemplo: Asesor de Código (conoce tu código)



IA generativa

Preparación de datos para Asesor de Código





IA generativa

Consideraciones para adoptar un LLM en la industria

Consideración	Descripción
Selección Modelo LLM	El Modelo LLm Aplicable 1. Puro LLm 2. Valor agregado al LLm 3. Open Source LLm
Accesibilidad y Desarrollo	Opciones de desarrollo que hagan accesible al LLm para uso de negocios
Enfoque de Adaptacion	Adaptación de modelos para trabajar con datos de organizaciones para lograr objetivos de precisión y rendimiento.
Preparación empresarial	Garantizar que los modelos cumplan con las exigencias de seguridad, confiabilidad, interoperabilidad y responsabilidad de la empresa.
LLm Ops	modelo operativo, procesos, marcos para producir y monitorear LLms para la empresa

IA generativa

Ejemplo de creación LLms comprimido OpenSource con RAG

Utilización de Modelo Open Source, del tipo meta Llama pre-entrenado como LLM y backend de integración para un asistente de investigación local basado en RAG.
Utilizando fuente de información web (wikipedia) y datos locales como Pds, Txt, etc

```python
from llama_cpp import Llama

# Cargar el modelo preentrenado
modelo = Llama.load("https://spanishchat-7b.Q4_K_M.gguf")

# Cargar los documentos y las páginas web
documents = ["Documento1.txt", "Documento2.txt", "Documento3.txt"]
web_pages = ["https://en.wikipedia.org/wiki/Fender_Stratocaster",
             "https://en.wikipedia.org/wiki/Gibson_Les_Paul",
             "https://en.wikipedia.org/wiki/Guitar_amplifier" ]

# Procesar los documentos y las páginas web
processed_docs = modelo.process_docs(documents)
processed_web_pages = modelo.process_web_pages(web_pages)

# Realizar una búsqueda de texto en los documentos y las páginas web
query = "texto de búsqueda"
results = modelo.search(processed_docs, processed_web_pages, query)

# Imprimir los resultados
for result in results:
    print(result)
```

IA generativa en la industria

UAI
Universidad Abierta Interamericana

¡¡Muchas Gracias!!

¿Preguntas?

Ing. Maximiliano Bonaccorsi