Aplicaciones con LLMs

Inteligencia Artificial e Ingeniería del Conocimiento

Constantino Antonio García Martínez

Universidad San Pablo Ceu

APIs Vs. Modelos Locales

LLM APIs vs. Modelos Locales

APIs Comerciales

Ventajas:

- Sin requisitos de hardware
- Modelos de última generación
- Escalabilidad instantánea
- Mantenimiento por el proveedor
- Actualizaciones frecuentes

Desventajas:

- Coste por uso
- Dependencia de conexión
- Datos enviados a terceros
- Límites de rate-limit
- Vendor lock-in

Modelos Locales

Ventajas:

- Control total de datos
- Sin costes recurrentes
- Privacidad garantizada
- Sin límites de uso
- Funciona offline

Desventajas:

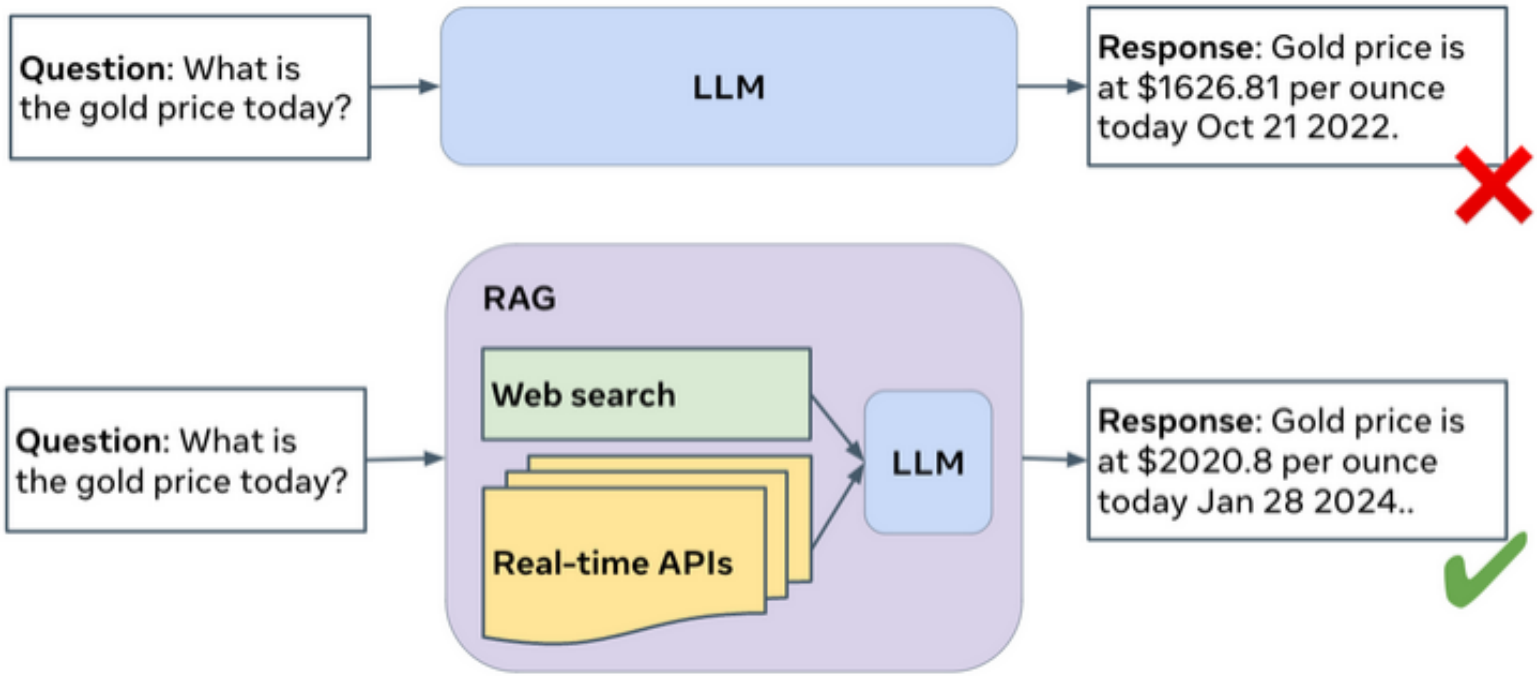
- Requiere hardware potente
- Mantenimiento propio
- Modelos menos capaces
- Complejidad de deployment
- Actualizaciones manuales

Acceso a LLMs por API: Demo

Example: 7-api_demo_mistral.py

Caso de Estudio: RAG

RAG: Retrieval-Augmented Generation



[1]

[1] Fuente: https://kddcup24.github.io/index.html

RAG: Retrieval-Augmented Generation

Example: 8-RAG.ipynb

Vector Databases

Vector Stores: Gestión Eficiente de Embeddings

¿Por qué necesitamos vector stores especializados?

- **Escala**: Una base documental puede generar millones de embeddings
  - 10,000 documentos $\times$ 100 chunks/doc = 1M embeddings
  - Vector de 768 dimensiones $\times$ 4 bytes = 3KB por embedding
  - Total: $\sim$3GB solo en embeddings
- **Búsqueda eficiente**: Calcular similitud con 1M vectores es costoso
  - Búsqueda lineal: $O(n \cdot d)$ donde d es la dimensionalidad
  - Vector stores usan índices aproximados (ANN: Approximate Nearest Neighbor): $O(\log n)$
  - Algoritmos:
    - Hierarchical Navigable Small World graphs (HNSW)
    - Inverted File Index (IVF)
    - Locality-Sensitive Hashing (LSH)
- **Persistencia y actualización**: Añadir/eliminar documentos dinámicamente
- **Filtrado y metadatos**: Búsqueda híbrida (semántica + metadatos)

Algunos de los Vector Databases más usados son: Pinecone, Chroma, FAISS, Weaviate, Milvus, Qdrant, etc.