

Plataformas y herramientas

Hugging Face

Hugging Face ofrece el **Model Hub**, repositorio público con millones de modelos preentrenados (textuales, vision, multimodales) y conjuntos de datos. También proporciona **Spaces**, un directorio de aplicaciones de IA interactivas compartidas. La biblioteca *Transformers* de HF permite cargar, ajustar o servir estos modelos bajo PyTorch o TensorFlow. Por ejemplo, hay APIs para clasificación de texto, traducción, generación de imágenes, etc. HF dispone de planes gratuitos (por ejemplo: almacenar modelos públicos gratis) y servicios de inferencia pagos (~\$0.033/hora por endpoint dedicado), lo cual se detalla en su página de precios. Esta plataforma es muy usada en investigación y educación ya que facilita acceder a modelos sin grandes recursos propios.

LangChain

LangChain es un framework de código abierto (Python/JavaScript) diseñado para crear aplicaciones que combinan LLMs con fuentes de datos externas. Permite “encadenar” pasos: por ejemplo, recibe la consulta del usuario, la procesa o traduce si es necesario, luego realiza una búsqueda o recuperación de información relevante en bases de datos o documentos, y finalmente alimenta al LLM con ese contexto para generar una respuesta informada. Así, LangChain facilita construir agentes de IA o chatbots especializados (ej. un chatbot médico que primero busca en un repositorio de investigaciones antes de responder). Es modular y extensible, y su propósito educativo es mostrar cómo integrar datos reales en aplicaciones basadas en LLMs.

Google Colab

Google Colab es un servicio de **notebooks Jupyter en la nube** (no hay que instalar nada local) que combina código ejecutable, texto enriquecido e imágenes en un mismo documento. Proporciona un entorno “listo para usar” con Python y bibliotecas de ML preinstaladas, facilitando la colaboración entre usuarios (compartición de notebooks vía Google Drive). Una ventaja clave es que Colab regala cómputo acelerado: acceso gratuito a GPUs (por ejemplo NVIDIA Tesla T4) y TPUs (Google’s Tensor Processing Units). Esto permite entrenar o inferir modelos sin invertir en hardware costoso. Colab tiene límites de uso (tiempo de sesión, RAM) en el plan gratuito, y ofrece

versiones *Pro* de pago para mayor capacidad. Por su sencillez, es muy popular en educación científica, permitiendo a profesores y alumnos ejecutar experimentos de IA sin infraestructura local.

Colab | Google for Developers

Figura: Interfaz conceptual de Google Colab con asistencia de IA. Colab combina código ejecutable y texto en notebooks colaborativos accesibles desde el navegador. Permite compartir trabajo con estudiantes y tiene asistentes de IA para autocompletar código.

Kaggle Notebooks

Kaggle (plataforma de ciencia de datos de Google) también ofrece notebooks colaborativos en la nube. Al igual que Colab, proveen GPUs gratuitas (ej. NVIDIA Tesla P100) y un entorno de Python listo, pero están más orientados a análisis de datos (con acceso directo a datasets de la comunidad) y a la competencia en proyectos de ciencia de datos. Son gratuitos y fáciles de usar para docentes que guían prácticas de análisis de datos con IA.

AWS SageMaker y Azure ML

Son **servicios en la nube** de AWS (Amazon) y Microsoft respectivamente. Permiten el entrenamiento, ajuste, despliegue y monitorización de modelos de ML a escala industrial. Por ejemplo, Azure Machine Learning es “un servicio en la nube que acelera y administra el ciclo de vida del aprendizaje automático”. Ofrece notebooks integrados, interfaces gráficas (Designer) y soporte para frameworks abiertos (PyTorch, TensorFlow, Scikit-learn). SageMaker similarmente. Ambos requieren pago según recursos usados (CPUs/GPUs en la nube), por lo que su uso educativo suele limitarse a instituciones con presupuesto o al plan gratuito inicial. Su ventaja es la escalabilidad y la integración con servicios empresariales (bases de datos, web services), aunque a costa de mayor complejidad y costo.

Ollama y LM Studio

Estas son **herramientas locales** para correr modelos de lenguaje en PCs sin depender de nube. *Ollama* (CLI/API) e *LM Studio* (aplicación de escritorio) permiten descargar y ejecutar LLMs open-source en Windows, Mac o Linux. Son gratuitas para uso personal. Por ejemplo, LM Studio anuncia: “Run AI models, locally and privately”, soportando modelos como gpt-oss, Qwen3 o Gemma3 en tu propio hardware. La figura ilustra su interfaz tipo IDE para inferir con un modelo en local. Al usarlos, los datos permanecen en tu dispositivo, lo que protege la privacidad. Estas plataformas

son útiles en entornos educativos donde no se desea o no se puede usar la nube.

LM Studio - Local AI on your computer

Figura: Captura esquemática de LM Studio ejecutando un modelo de IA localmente. LM Studio es una app gratuita que permite correr LLMs en el ordenador, incluyendo asistentes de autocompletado. Al usarse localmente, mantiene los datos del usuario fuera de la nube.

vLLM

vLLM es una biblioteca de inferencia desarrollada inicialmente en la Universidad de Berkeley (ahora con soporte de NVIDIA) diseñada para servir LLMs con **gran rendimiento y eficiencia de memoria**. Permite procesar muchas solicitudes en paralelo, optimiza la gestión de memoria de las llaves-valor de atención, y utiliza técnicas de cuantización (INT4/INT8, FP8). Se integra con modelos de Hugging Face y soporta GPUs (NVIDIA, AMD, Intel) e incluso CPUs, y hasta TPUs mediante plugins. *vLLM* ofrece un servidor API compatible con OpenAI para inferencia, ideal para montar demos educativas de IA de forma eficiente.

Bibliotecas principales

- **PyTorch:** Es una *biblioteca de tensores* optimizada para aprendizaje profundo sobre GPUs/CPUs. Ofrece un modo de ejecución ansioso (fácil de usar en investigación) y herramientas para producción (TorchScript, TorchServe). PyTorch cuenta con un ecosistema rico (TorchVision, TorchAudio) para visión, audio y NLP, y está soportado por principales nubes (AWS, GCP, Azure).
- **TensorFlow:** Es una plataforma de Google de extremo a extremo para ML. La página oficial destaca que “TensorFlow facilita la creación de modelos de AA que pueden ejecutarse en cualquier entorno”. Posee API intuitivas (como Keras) y bibliotecas auxiliares (TFX, TensorFlow Lite para móviles, etc.). TensorFlow también permite exportar modelos a otros formatos (SavedModel, TFLite) para desplegarlos en servidores o dispositivos móviles.
- **Scikit-learn:** Biblioteca de Python de propósito general para aprendizaje automático clásico. Brinda implementaciones eficientes de algoritmos de regresión, clasificación, clustering y reducción de dimensionalidad. Como señala IBM, “*Scikit-learn es una de las bibliotecas de machine learning más utilizadas*” en Python. Aunque no está diseñada para redes neuronales profundas, es indispensable para tareas de ciencia de datos (preprocesamiento, validación cruzada, métricas) en educación.



Updated 2026-03-04 17:44:52 CET by Luis Hueso